Aplicação de técnicas de *data mining* ao setor vinícola

por

Rui Nuno Ribeiro de Souza Roza

Projecto de tese de Mestrado em

Modelação, Análise de Dados e Sistemas de Apoio à Decisão

2016

Orientador: Professor Doutor Pavel Brazdil (FEP)

Co-orientador: Professor Doutor Óscar Felgueiras (FCUP)

Breve Nota Biográfica

Rui Nuno Ribeiro de Souza Roza, natural do Porto, licenciado em Administração e Gestão de Marketing pelo IPAM – Instituto Português de Administração de Marketing.

A maior parte dos vinte anos do percurso profissional foi no setor bancário, quase sempre ligado ao tratamento, gestão de acessos informáticos, arquivo, tanto físico como electrónico, e auditoria, por forma a assegurar a máxima qualidade da informação estratégica dos clientes.

Tendo começado pela área de Telemarketing e da Gestão de Acessos Informáticos foi destacado para um banco do grupo em Moçambique, onde permaneceu oito anos, implementando medidas de detecção, análise, correcção e prevenção da informação estratégica do cliente, assim como, assegurando a gestão do arquivo, tanto físico como digital, das fichas de informação pessoais dos clientes. Em França, permaneceu num banco do mesmo grupo durante 4 meses, desenvolvendo um projeto de identificação e correção dos registos duplicados existentes na base de dados estratégica de clientes. De regresso a Portugal, integrou uma equipa de gestão de risco de crédito empresas onde procedeu à actualização dos códigos de actividade de económica de todos os registos existentes na base de dados estratégica de clientes.

Por todo este percurso profissional, a elaboração do mestrado nesta área justificou-se pela procura de melhorar os conhecimentos sobre o que as novas tecnologias tem ao nosso dispor para melhorar as tomadas de decisão da empresa na senda da prestação de um serviço ao cliente de excelência.

Agradecimentos

O desenvolvimento e concretização desta dissertação não teria sido exequível sem o apoio, incentivo e disponibilidade de diversas pessoas que me acompanharam e contribuiram direta ou indiretamente ao longo destes meses, para atingir este objetivo. A todos, de um modo muito especial, o meu muito Obrigado.

Ao Professor Doutor Pavel Brazdil, pela ideia inicial do tema da tese; agradeço ainda pela disponibilidade, transmissão de conhecimentos, empenho, minúcia, compreensão e amizade, além de motivação contínua e humildade que, com toda uma imensa experiência detida ao longo dos diversos anos de investigação, me incentivou a desenvolver nos dois estudos agora aqui apresentados. Com o Professor Pavel aprendi também a ser mais pessoa.
Ao Professor Doutor Óscar Felgueiras, pela amizade, apoio e disponibilidade, mesmo nas horas mais "apertadas", para encontrar alguma solução viável.

Ao Dr. José Luís Reis, da Comissão de Viticultura da Região de Vinhos Verdes, pela disponibilidade, sugestão de ideias e apoio determinante para o estudo efectuado sobre os vinhos verdes. Desta mesma Comissão, agradeço também ao Dr. Paulo Martins pela ajuda na disponibilização dos ficheiros, interpretação da informação e revisão do relatório final.

Ao Dr. António Graça, da Sogrape, uma palavra de apreço pelas sugestões levantadas, disponibilidade revelada e esclarecimentos prestados nas diversas reuniões mantidas para a concretização da ideia do estudo sobre o painel de provadores. Não posso deixar também de agradecer aqui tanto a participação nas reuniões como a ajuda prestada, quer na revisão das ideias quer na extração da informação, pela Dra. Natacha Fontes e pela Dra. Joana Martins.

A todos os meus Amigos pela ajuda e apoio e, em particular, à Sónia Teixeira, ao Dewan Fayzur e ao Rui Nunes pelo companheirismo, amizade, disponibilidade e esclarecimento de algumas dúvidas importantes que permitiram ultrapassar os problemas que foram surgindo ao longo deste estudo. Dewan Fayuzur, um especial obrigado pela companhia e pelas longas conversas nos nossos almoços, momentos importantes de descontração, também necessários, e de discussão de ideias.

E por último e, sem nunca ser de menos...

Obrigado, À Sandra, pela tua luz de presença, alegria, apoio incondicional, assim como, pela confiança depositada em mim, desde o primeiro momento, para que este projeto se concretizasse. À Catarina, minha filha, pela tua energia que faz mover montanhas, pelo teu sorriso aberto e pela vontade de conhecer o mundo e por te espantares com o simples e com tudo o que é novo. À minha mãe, imagem de determinação que tenho sempre presente, vai toda a minha admiração e carinho. Obrigada por acreditar em mim! E à memória de meu pai, pela sua presença em espírito, vai toda a saudade dos momentos que passamos e dos que poderíamos ter passado. Obrigado pela força que continuo a sentir.

Resumo

Nas últimas décadas, tanto o setor vinícola como as tecnologias de informação têm sido alvo de grandes desenvolvimentos. No caso do primeiro, ocorreram mudanças radicais, desde o espaçamento da vinha à modernização das infra-estruturas, do controlo de qualidade e do tratamento dos solos à adaptação climática. No segundo, a introdução de novos conceitos como o business intelligence, que engloba novas técnicas, tais como, *data mining* e *data warehouse*, permitiu recolher, armazenar e tratar, quantitativa e qualitativamente, a informação de modo aos decisores melhorarem as suas tomadas de decisão.

Tendo presente estes princípios, o atual trabalho, composto por dois estudos, tem como objetivo abordar alguns dos pressupostos que interligam o setor vinícola com técnicas específicas de *data mining* e verificar até que ponto esta interligação pode melhorar a tomada de decisão.

O primeiro estudo, efectudo com uma entidade reguladora, a *Comissão de Viticultura da Região dos Vinhos Verdes (CVRVV)*, pretendeu, através da conjugação da informação existente nas análises físico-químicas e organoléticas e utilizando técnicas de *data mining*, agrupar por *clusters* de vinhos semelhantes para encontrar as características diferenciadoras entre os vinhos das nove sub-regiões existentes na Região Demarcada de Vinhos Verdes. Este estudo permitiu identificar agrupamentos (*clusters)* com características bem identificadas e referindo-se a sub-regiões específicas.

O segundo estudo, efectuado com a *Sograpе Vinhos S.A.*, consistiu em avaliar um painel de provadores, a partir duma base de dados de provas de vinho, efectuando a análise das notas atribuídas por esses provadores. Pretendeu-se verificar se uns se destacam por um *viés* significativo, positivo ou negativo, em relação aos outros provadores. Nos casos em que isso se verificou, pretendeu-se ainda encontrar uma caracterização das condições em que um dado provador sobre- ou sub-avaliou certos vinhos. Além disso, procuramos ainda identificar as características de vinho mais importantes que condicionaram a sua decisão. Essa análise incidiu sobre as regressões lineares geradas por *model trees*, implementados sob o nome M5P em Weka. O nosso trabalho pode ser usado pela empresa para ou restruturar o painel de avaliadores e/ou para procurar formas de corrigir o enviesamento observado.

Abstract

In last few decades both the wine sector as information technology have undergone major developments. Regards the former, there were radical changes from the spacing of the vines to the modernization of infrastructure, quality control and soil treatment prompted by climate change.

Regards the second item referred to, the introduction of new concepts such as business intelligence, which includes new techniques such as *data mining* and *data warehousing*, permitted to collect, store and process quantitatively and qualitatively, the information so that the decision makers could improve their decision making.

Bearing these principles in mind, this paper consisting of two studies, aims to address some of the links between the wine sector and specific techniques of data mining and verify to what extent this interconnection can improve decision making.

The aim of the first study, carried out in collaboration with a regulatory body, the *Viticulture Commission of the Region of Vinho Verde (CVRVV)*, was to group into clusters with similar physic-chemical and organoleptic characteristics the wines of nine sub-regions in the demarcated region of *vinho verde*. This study identified clusters (clusters) with well-identified characteristics, some of which referring to specific sub-region(s).

The aim of second study, carried out in collaboration with *Sograpе S.A.*, was to evaluate a panel of wine tasters, using a given wine tasting database and in particular, the scores assigned by these tasters. Our aim was to verify whether some of the wine tasters exhibit a significant bias, positive or negative, in relation to the other tasters. In cases when such bias was identified, our objective was to characterize the conditions under which a particular taster over- or under-evaluated certain wines. In addition, we sought to identify the most important wine characteristics that affected his decision. This analysis focused on the linear regression model generated by *model trees*, implemented under the name of *M5P* in Weka software. Our work can thus be used by the company to restructure the evaluation panel and/or alternatively to look for ways to correct the observed bias.

Índice

Índice das Figuras

Índice das Tabelas

Abreviaturas

CVRVV – Comissão de Viticultura da Região de Vinhos Verdes

DO – Denominação de Origem

DR – Diário da República

SVM – Máquinas de vetores de suporte

NN – Redes neuronais

1. Introdução - motivação e objetivos

Nas últimas décadas, tanto o setor vinícola como as tecnologias de informação sofreram grandes desenvolvimentos. No caso do primeiro, ocorreram mudanças radicais, desde o espaçamento da vinha à modernização das infra-estruturas, do controlo de qualidade e do tratamento dos solos à adaptação climática (Silva, 2015). No segundo, o conceito de *business intelligence*, que engloba novas técnicas, tais como, *data mining* e *data warehouse*, permitiu recolher, armazenar e tratar, quantitativa e qualitativamente, a informação de modo aos decisores melhorarem as suas tomadas de decisão (Braga, 2009).

Tendo presente estes princípios, o atual trabalho, composto por dois estudos, tem como objetivo abordar alguns dos pressupostos que interligam o setor vinícola com técnicas específicas de *data mining* e verificar até que ponto esta interligação pode melhorar a tomada de decisão.

O primeiro estudo, resultante do contacto com uma entidade reguladora, a Comissão de Viticultura da Região dos Vinhos Verdes (CVRVV), pretende, através da conjugação da informação existente nas análises físico-químicas e organoléticas e utilizando técnicas de *data mining*, agrupar por *clusters* de vinhos semelhantes para encontrar as características diferenciadoras entre os vinhos das nove sub-regiões existentes na Região Demarcada de Vinhos Verdes. Para isso, foi efectuado, numa primeira fase de pré-processamento, a análise e tratamento do ficheiro inicial, do qual resultaram para o estudo 4.941 amostras de vinho verde branco, seguindo depois para a fase de agrupamento (*clustering*). Na última fase foram definidos os grupos (*clusters*) e as suas caracterizações, a serem validados por técnicos da CVRVV.

O segundo estudo, surgido após contacto com a empresa produtora de vinhos Sogrape Vinhos S.A., consiste em avaliar o comportamento de um painel de provadores, a partir duma base de dados de vinhos, com informação recolhida ao longo de oito anos (2007-2015), efectuando a análise das características organolépticas das amostras de vinho sujeitas a classificação por esse mesmo painel de provadores, composto por 15 enólogos. Estes especialistas classificam o vinho com um valor, numa escala de pontuação entre 0 e 20, tendo por base um vinho de referência. Para isso, e após uma primeira fase de pré-processamento da qual resultaram para o estudo 335 provas de vinho branco, foi efectuada uma análise estatística que permitisse identificar o comportamento dos provadores de modo a identificar os que atribuem notas acima (ou abaixo) da média. Seguidamente, pretendeu-se perceber quando é que um determinado provador aplicava a sua regra diferente do geral e se a forma da regra usada por esse provador é comparável com a regra geral dos restantes provadores sendo para isso,

usadas *model trees*, implementados sob o nome M5P, no *software* Weka e a regressão linear simples.

A presente tese apresenta no capítulo 2 o estado da arte na área da vinicultura, identificando alguns estudos e projectos que interligam as análises, tanto físico-químicas como organolépticas, com as técnicas de *data mining*, de modo a melhorar os processos de tomada de decisão.

No capítulo 3 são identificadas as técnicas a utilizar nos dois estudos apresentados, ou seja, para o primeiro, o método de *clustering*, para a identificação das similaridades entre os vinhos das nove sub-regiões de vinho verde branco, e para o segundo, a utilização da regressão linear e de árvores de regressão M5P, para avaliar o comportamento do painel de provadores.

O capítulo 4 aborda o primeiro estudo, da CVRVV, sobre a semelhança entre os vinhos verdes brancos existentes nas nove sub-regiões.

O capítulo 5 apresenta o segundo estudo, da Sogrape, que analisa um painel composto por 15 provadores de vinhos de modo a reconhecer os provadores que apresentam maior viés, positivo ou negativo, em relação à média e se há algumas características físico-químicas ou organolépticas que poderão influenciar a sua decisão na atribuição da classificação às amostras de vinho.

Por último, no capítulo 6 são referidas as conclusões aos dois estudos, assim como, eventuais futuros estudos que poderão, no caso da CVRVV, elaborar também o mesmo para os vinhos verdes tintos e estudar as relações de agrupamento por forma a verificar se estas têm uma evolução temporal. No caso da Sogrape, os desenvolvimentos podem ser usados para corrigir o viés, positivo ou negativo, das classificações atribuídas pelos provadores às amostras de vinho e/ou eliminar os provadores que não satisfazem os critérios exigidos.

2. Estado da Arte

2.1. Diferentes trabalhos na área da vinicultura

Nesta área foram já realizados alguns trabalhos, uns comparando características físico-químicas das amostras de vinhos, ou seja, objetivas, com as características organolépticas, de caráter mais subjectivo, tais como, cor, limpidez, aroma e sabor, e outros conjugando os comentários dos provadores de vinhos com os localizações geográficas, o clima e as regiões onde se encontram as castas de melhor qualidade.

2.1.1. Previsão da qualidade do vinho a partir de dados físico-químicos

Sendo Portugal um país em que o setor vinícola tem bastante expressividade na economia, tendo vindo a produzir anualmente cerca de 6.000.000 hl desde 2012 e, de acordo com a Figura 1, representando 49,1% do total do volume de vendas das indústrias das bebidas, mantendo nos últimos anos a mesma média de produção, conforme ilustra a Figura 2, o factor qualidade desempenha um papel fundamental (INE, 2015).

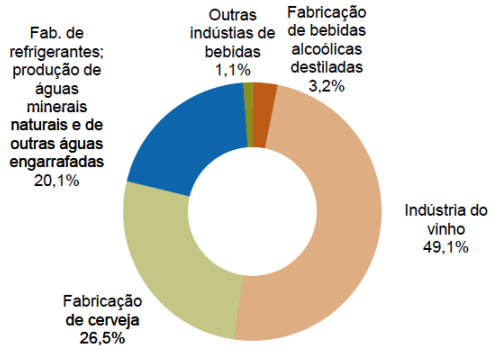

Figura 1 - Valor de vendas das Indústrias das Bebidas 2013 (INE, 2015)

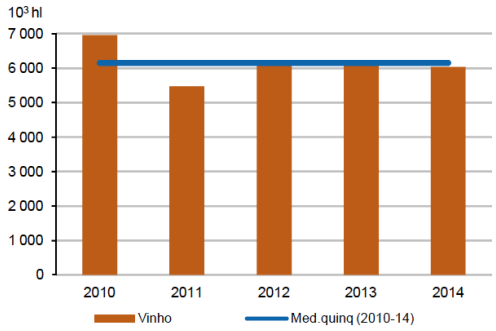

Figura 2 - Produção de Vinho (INE, 2015)

As *características físico-químicas* (perfil analítico – título alcoométrico e acidez fixa) são um dos elementos fundamentais para classificar o tipo de vinho e determinar as suas características, existindo instituições públicas que procedem ao controlo dos limites inferiores e superiores destes dados analíticos e posterior certificação (CVRVV, 2014). No caso concreto dos vinhos verdes, em que é a CVRVV que lidera o processo, é aposto um selo de garantia que certifica a "DO" e a "Indicação Geográfica", neste caso concreto, Vinho Verde, para o primeiro e Minho, para o segundo (CVRVV, 2014).

Para além destas características, temos os *atributos organolépticos*, ou seja, *cor*, *limpidez*, *aroma* e *sabor*, cujos critérios são mais subjectivos (Ribeiro et al., 2009a), comprovados através de prova sensorial por enólogos de modo a garantir a genuinidade e qualidade dos vinhos (CVRVV, 2002). Cortez et al. (2009a) criaram modelos capazes de prever as características organolépticas - *cor*, *limpidez*, *aroma*, *sabor* - mais subjectivas, através das características físico-químicas, mais objetivas, de modo a prever as preferências vinícolas apontadas, através das avaliações do enólogo, e melhorar a produção de vinho.

Segundo Grainger (2009), quanto melhor um vinho conseguir expressar e sobressair as suas características próprias - físico-químicas e organolépticas - maior será a sua qualidade. Os vinhos são produzidos em termos de:
- variedade, sendo que quanto mais expressiva for a sua variedade maior a qualidade;
- frutado, ou seja, quanto mais aroma a frutas primárias exalarem mais qualidade tem;
- geografia/região, quanto mais representativo das características da região (ex: acidez, adocicado, entre outros) também melhor será a sua qualidade.

Usualmente, os laboratórios fazem testes às características físico-químicas, tais como, a densidade, o álcool e o pH (Cortez et al., 2009b). Para prever a qualidade do vinho, através de dados físico-químicos, têm sido usadas técnicas de *data mining*, tais como, regressão linear múltipla, classificadores de tipo, máquinas de vectores de suporte (SVM) e redes neuronais (NN), dos quais estes últimos têm ganho mais preponderância.

Os autores acima citados efectuaram um estudo com um *dataset*, constituído por 4.898 amostras de vinho verde branco, com as características analíticas – dados físico-químicos – e as preferências vinícolas, definindo uma escala, de 0 (muito mau) a 10 (muito bom), para serem utilizadas na previsão das preferências vinícolas do consumidor, aplicando as técnicas de *data mining* referidas no parágrafo anterior. Verificaram que o SVM atingia os melhores resultados, atingindo uma precisão que variava entre 64,3% e 86,8%, muito melhor do que os resultados de um classificador aleatório, pois o *dataset* continha 6 a 7 classes na escala de qualidade (valores entre 3 e 8 a 9). Foi sugerido que os resultados obtidos poderão ajudar nas avaliações efectuadas pelo enólogo e melhorar na produção e venda dos produtos (Cortez et al., 2009b). Se, eventualmente, os valores registados pelo modelo predictivo das preferências vinícolas forem muito diferentes dos registados pelo enólogo, este último poderá repetir o teste para confirmar.

Como conclusão verificaram que, como as avaliações do perfil sensorial são mais subjectivas e os resultados obtidos através de uso de classificadores são baseados em testes

mais objetivos, este tipo de solução poderia ser utilizada para, de futuro, melhorar a rapidez e qualidade do trabalho do enólogo no suporte à tomada de decisão. Para além disso, esta abordagem poderia também ser utilizada no apoio à formação de novos enólogos (Cortez et al., 2009b).

2.1.2. Modelos de crescimento de vinha: Previsão da quantidade e da qualidade da uva

Segundo Dougherty (2012), a geografia desempenha um papel preponderante na produção de vinho e no processo de crescimento da vinha, tanto em termos de quantidade como de qualidade – gosto e aroma único que dependem de região para região, de *terroir*[1] para *terroir* e das técnicas utilizadas. O sucesso de um bom vinho não depende só de um produto de qualidade, mas também da experiência adquirida na área de pesquisa e desenvolvimento – R & D – da própria empresa e de uma abordagem inovadora à produção de vinho, marketing e vendas (Parliament of Australia, 2001). Alguns autores (Santos et al., 2013; Daux et al., 2012; Gouveia et al., 2011) defendem que o clima exerce um papel fundamental na produtividade da vinha, dado que esta é muito sensível às diferenças climatéricas, nomeadamente, temperatura do ar e precipitação. Segundo Santos et al. (2013) o reconhecimento de fortes relações entre os fatores atmosféricos e as características da vinha são muito importantes para a análise de cada região vinícola.

No caso concreto deste estudo sobre a produção vinícola no vale do Douro, Santos et al. (2013) utilizaram a *regressão linear multivariada* aplicada à produção vinícola de uma longa série temporal, entre 1932 e 2010, para modelar o crescimento. Os autores, identificaram que a forte pluviosidade e temperaturas baixas na brotação e inflorescência, que ocorre nos meses de fevereiro/março, e as temperaturas quentes durante a floração e o desenvolvimento da uva, no mês de maio, são favoráveis a uma produção elevada. Segundo um estudo no vale do Douro, por Gouveia et al. (2011), "*um bom ano de produção reflecte a alta actividade fotossintética na Primavera seguida de reduzida verdura durante o Verão. Pouca precipitação durante o estágio de crescimento, em Março, tem um efeito positivo e altas temperaturas, durante o fim da Primavera, são benéficas para a produção*". Para isso levou a efeito dois modelos de regressão linear com o propósito de estimar a produção vinícola do Douro nos princípios de março e na estação média - julho - do ciclo vegetativo da vinha. Os autores (2011), apresentam a produção vinícola, reproduzida na Figura 3, utilizando um modelo de regressão linear no

[1] Segundo (2012, Unwin, T.) há muito tempo que os produtores vinícolas franceses usam o termo *terroir* quando se referem à complexa interacção entre todos os aspectos físicos da geologia, dos solos, do clima, geomorfológicos e de vegetação combinados para criar um local particular onde as uvas se desenvolvem.

início da estação, e na Figura 4, utilizando um modelo de regressão linear no meio da estação, havendo um único *outlier* a registar, em 2004:

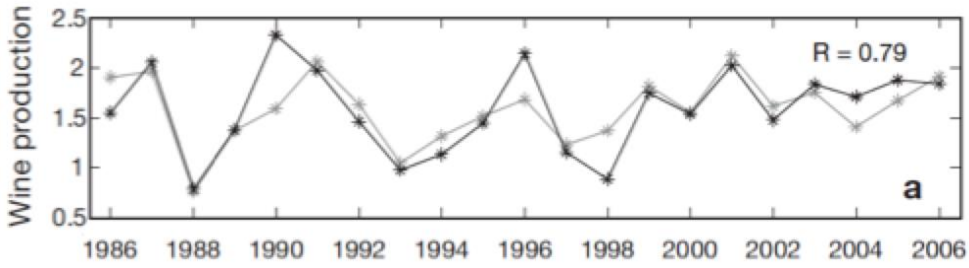

Figura 3 - Produção vinícola observada na região do Douro e valores correspondentes previstos pelo modelo no início da estação (curva cinzenta escura)

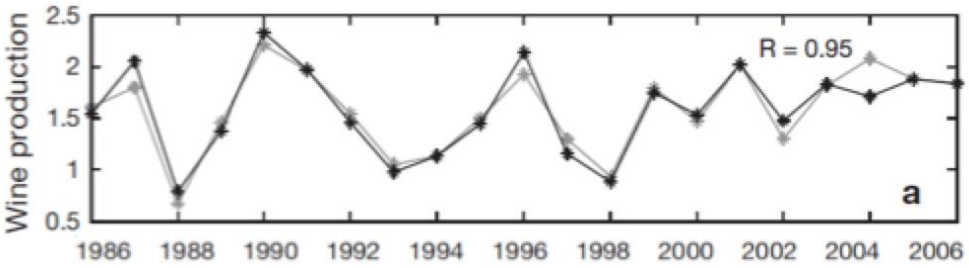

Figura 4 - Produção vinícola observada na região do Douro e valores correspondentes previstos pelo modelo no meio da estação (curva cinzenta escura)

Como se pode verificar, a Figura 4 apresenta melhores resultados em termos de medida R, qualidade de ajustamento (*fit*), provavelmente porque no meio da estação o sistema tem mais informação disponível.

2.1.3. Utilização de *Data Mining/Machine Learning* no controlo de produção de vinho

De acordo com Ribeiro et al. (2009b) o processo de vinificação é um dos que mais contribui para a qualidade do vinho, onde se transforma a uva em vinho, e é composto pelas seguintes fases (CVRVV, 2014):

1 - Esmagamento da uva; 2 - Prensagem; 3 - Decantação; 4 - Fermentação (mosto – líquido das uvas – separado do engaço e das películas) e 5 - Análise laboratorial.

No caso do vinho tinto, como a casca da uva dá a coloração ao vinho, a fermentação é feita em presença das partes sólidas da uva. Segundo Grainger et al. (2005), a sua importância está no facto de na polpa existirem açúcares (fructose e glucose) que, durante a fermentação, por acção das enzimas da levedura, se transformam em alcool etílico + dióxido de carbono + calor, para além de outros produtos. O enólogo tem que controlar todo este processo de modo a produzir um vinho saboroso, balanceado e com o seu estilo próprio.

Para os autores Ribeiro et al. (2009b), este processo é, tradicionalmente, desenvolvido por enólogos que analisam os atributos organolépticos, tais como, *cor*, *espuma*, *sabor* e *cheiro*, fundamentais para a produção de vinho e para a divulgação do marketing associado. Num estudo apresentado por estes autores foram utilizadas duas técnicas de *data mining*/*machine learning* - árvore de decisão e regressão linear, para atingir o objetivo da classificação, ou seja, representar a relação entre os atributos químicos e a variável que representa a classe. Segundo Mitchell (1997), a árvore de decisão é um dos métodos mais usados na inferência indutiva. Este método permite aproximar funções discretas a dados com ruído e é capaz de aprender expressões disjuntivas. As árvores de decisão classificam as instâncias ordenando-as de modo a formar uma árvore, desde a raíz se separar em todos os nós folha, que possuem a classificação de cada instância. Na árvore, cada nó especifica um teste de algum atributo da instância e cada ramo que descende desse nó diz respeito a um dos possíveis valores desse atributo.

Os dados do estudo foram recolhidos durante a fase de produção do vinho, por um período de 4 anos, numa empresa vinícola da região do Minho produtora de vinho verde tinto, e possuíam dois tipos de atributos: caraterísticas químicas (objetivas) e organolépticas (subjectivas). Os dados foram repartidos por: dados subjectivos, com duas classes, *média* e *bom*; dados físico-químicos (objetivos) em intervalos de 5 valores. Foi construída uma árvore de decisão do atributo *cor do vinho*, correspondente à representação de um conjunto de regras que segue uma hierarquia de atributos, expressando uma lógica condicional (Ribeiro et al., 2009b), conforme Figura 5.

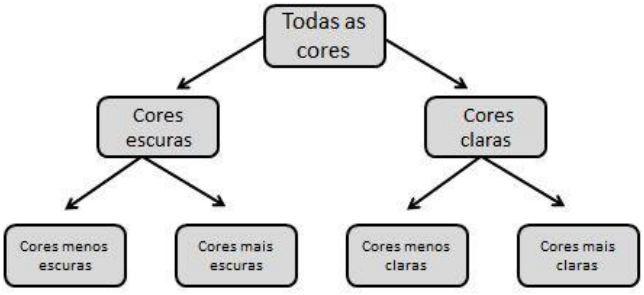

Figura 5 - Árvore de decisão do atributo cor do vinho

Os autores conseguiram obter uma precisão entre 85% e 98%, demonstrando que os modelos de *data mining/machine learning* podem ser utilizados para prever atributos subjetivos no processo de vinificação, com base em parâmetros químicos.

2.1.4. Utilização de técnicas de *data mining* na análise dos comentários dos provadores

Sallis et al. (2008) efectuaram um estudo, em diversas regiões da Nova Zelândia, pretendendo identificar as relações existentes entre as castas, suas condições de crescimento e sua

localização, utilizando o SOM - *Kohonen self-organising map*, ou seja, um tipo de rede neural artificial, conjugado com a análise de componentes principais e as técnicas de *data mining: text mining* e *k-means*. Os comentários dos provadores, quanto à análise sensorial, foram extraídos por *text mining*, sendo depois utilizada a análise de componentes principais conjugada com o *clustering* (*k-means*). Estes comentários foram agrupados e cruzados com os dados geo-referenciados, utilizando o SOM, por forma a mapear as características sensoriais, os *terroirs* e as castas mais significativas por região.

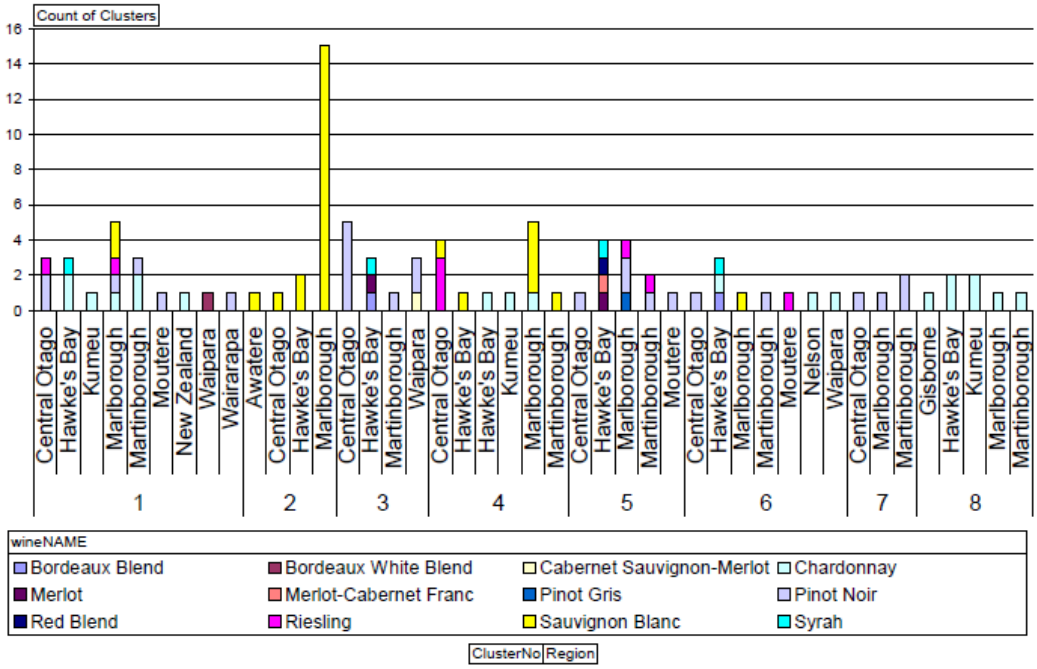

Figura 6 - Distribuição dos clusters de castas por regiões

O gráfico da Figura 6 ilustra o resultado final do estudo contendo a distribuição dos comentários dos provadores por 8 *clusters* SOM. Os *clusters* refletem os tipos de vinhos, como por exemplo, o *cluster* 2 tem os *Sauvignon Blanc* distribuídos pelas regiões Awatere, Central Otago, Hawke's Bay and Marlborough e o *cluster* 8 tem todas as castas *Syrah* em Gisborne, Hawke's Bay, Kumeu, Marlborough e em Martinborough.

3. Técnicas de *Data Mining* utilisadas nos nossos estudos

Os dois trabalhos a abordar nesta tese, sumariamente descritos no ponto 1.2., irão utilizar as técnicas de *data mining* - agrupamento (*clustering*) e - árvores de regressão, que abaixo são descritas.

3.1. Agrupamento (*Clustering*)

Clustering é um processo de agrupar um conjunto de objectos com determinadas características em vários grupos (*clusters*) de modo a que os objectos dentro de um *cluster* tenham alta similaridade (semelhança intra-*cluster*), mas os objectos de *clusters* diferentes (semelhança inter-*clusters*) tenham elevada dissimilaridade. Esta semelhança ou diferença é baseada na análise dos atributos dos objectos e, muitas vezes, também são utilizadas as medidas das distâncias entre estes – as mais comuns são a distância Euclideana e a distância de Manhattan para a sua identificação (Han et al., 2012; Gama, 2002). Esta técnica de *data mining* é de *aprendizagem não supervisionada*, pois não existem grupos pré-definidos, não se sabe nem o número de grupos e nem a estrutura dos mesmos (Gama, 2002). Segundo Mirkin (2005) os seus objetivos são:

(1) Estruturar, representando os objectos similares como conjuntos de grupos (*clusters*), com características idênticas entre si;

(2) Descrição dos *clusters* em termos de atributos, ou seja, quantas mais forem as características do objecto que satisfaçam a descrição dos *clusters* maior a possibilidade de aquele(s) pertencer(em) ao mesmo *cluster*;

(3) Associação, encontrar e maximizar inter-relações entre diferentes elementos do grupo, ou seja, padrões de alta similaridade inter-grupos;

(4) Generalização, fazendo afirmações gerais acerca dos dados e sobre as especificidades do assunto a que esses mesmos dados se referem;

(5) Visualização, representando as estruturas dos *clusters* como imagens visuais, tais como, gráficos de 1 dimensão - histograma, caixa de bigodes, entre outros; gráficos de 2 dimensões - gráfico de dispersão, entre outros (Mirkin, 2005; Han et al., 2012).

A técnica de *Clustering* tem diversos métodos, dos quais se destacam os métodos baseados em distâncias (*K-means*; *K-medoids*), de partição e os hierárquicos, que podem ser aglomerativos ou divisórios (Han et al., 2012; Gama, 2002). Conforme atrás referido, esta técnica é não supervisionada, e assim, não usa exemplos pré-classificados. Há, no entanto, algumas propriedades a considerar: Escalabilidade, em termos de espaço e tempo; Capacidade de trabalhar com diferentes tipos de dados – numéricos, categóricos, ordinais, entre outros e sensibilidade ao "ruído" e aos *outliers*. (Gama, 2002; Mirkin, 2005).

3.1.1. Algoritmo *K-Means*

O algoritmo *K-Means* é a principal técnica de *clustering* de partição e está presente na maior parte dos *softwares* (SPSS, SAS, Weka, Knime, entre outros). Este método é rápido, eficiente na memória e computacionalmente fácil tendo, no entanto, algumas susceptibilidades quanto à configuração inicial e quanto à estabilidade dos resultados (Mirkin, 2005). Segundo Gama (2002) este método escolhe, de forma aleatória ou não, K elementos diferentes para inicializar a identificação do centróide do grupo. No passo 2 é feita a associação de cada elemento ao centro mais próximo. E no passo 3 é recalculado o centro de cada grupo. Os passos 2 e 3 são repetidos até não serem encontradas mais alterações nos grupos. A maior dificuldade está na dimensionalidade, pois muitos algoritmos de *clustering* funcionam bem em bases de dados pequenas mas produzem resultados desiquilibrados quando se utilizam bases de dados de grandes dimensões.

3.1.2. Determinar o número ideal de *clusters*

Para determinar o número ideal de *clusters* a analisar foi utilizado o algoritmo *Simple EM* (*Expectation Maximisation*), que apesar de ser do *software* Knime é baseado no WEKA 3.7. Este algoritmo indica uma distribuição de probabilidade para cada instância que determina a sua probabilidade de pertença a cada um dos *clusters*. Deste modo, o algoritmo sugeriu criar 4 *clusters* por validação cruzada, efectuando os seguintes passos:

1. Colocando o número de *clusters* partindo de 1;
2. O conjunto de teste é dividido aleatoriamente em 10 partes;
3. O algoritmo EM é executado 10 vezes usando a divisão por 10 partes na forma de validação cruzada;
4. A log verossimilhança é a média de todos os 10 resultados;
5. Se a log verossimilhança aumentar, o número de *clusters* será incrementado em 1 e o programa prossegue para o passo 2.

Para uma melhor compreensão na Figura 7 são identificados os nós utilizados para a determinação do número ideal de *clusters*.

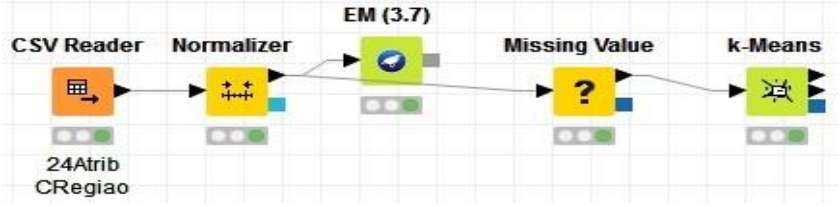

Figura 7 - Construção de conexões para obter os resultados, no Knime

3.2. Regressão Linear e árvores de regressão M5P

A técnica de regressão pode ser utilizada para modelar a relação entre uma ou mais variáveis independentes e a variável dependente ou de resposta que é de valores contínuos (Han et al., 2006; Gama et al., 2012). No contexto de *data mining*, as variáveis de previsão são os atributos de interesse que descrevem o caso, tornando-se no vector valores de atributos. Normalmente, os valores das variáveis são conhecidos (Han et al., 2006; Witten et al., 2011). Segundo Witten et al. (2011), os modelos lineares são mais fáceis de visualizar em duas dimensões, ou seja, o equivalente a desenhar uma linha reta através de um conjunto de pontos de dados, onde a linha recta indica a equação com a melhor previsão. De acordo com James et al. (2014) para cada observação do valor de previsão x_i (eixo das abcissas) – tal que i=1,2,…, n – existe um valor y_i de resposta associada sendo, por isso, apelidada de *aprendizagem supervisionada*. Pretende-se criar um modelo que dê resposta às variáveis de previsão, com o objetivo de prever com precisão a resposta para futuras observações ou melhor entender a relação entre as respostas observadas e as variáveis de *input*.

A equação pode ser traduzida conforme a Figura 8 (Han et al., 2006) sendo *X*, a variável independente, e *Y*, a variável de resposta, onde o *offset* de Y é representado pelo coeficiente b e a inclinação pelo coeficiente w_i:

$$Y = b + wX$$

Figura 8 - Equação de regressão linear simples

No caso concreto do setor vinícola e em alguns destes estudos, em particular, esta técnica irá permitir examinar as relações entre os diversos atributos do estudo e a variável objetivo, criando modelos de previsão a aplicar para futuras observações (James et al., 2014). Além da técnica descrita aqui existem diversos algoritmos que podem ser usados para os problemas de regressão, como por exemplo, *regressão logística*, *árvores de regressão*, *redes neuronais*, *SVM*s, *modelos aditivos*, entre vários outros (Brazdil, 2014).

No segundo estudo, referente à análise do comportamento dos provadores de vinhos, foram utilizados *model trees*, implementados sob o nome M5P no Weka, que combinam uma árvore de decisão convencional, que são sistemas hierárquicos com condições sequencialmente verificadas até chegar a um nó final (folha), com a possibilidade de incluir nestes últimos as funções de regressão linear. Segundo os autores (Frank et al., 2008), esta é uma técnica de previsão de valores numéricos contínuos.

4. Estudo de semelhança entre vinhos verdes brancos

Este estudo pretende decidir se o vinho de um certo tipo e sub-região se assemelha a outro doutra sub-região tendo, para isso, sido usada a técnica de agrupamento (*clustering*) e, em particular, o algoritmo *k-means*. A solução foi desenvolvida com o auxílio do Knime.

4.1. Descrição dos dados e Pré-processamento

O vinho verde, branco, tinto ou rosado, apresenta características próprias, definidas por lei (Dec-Lei nº 10/92), que variam de sub-região para sub-região, dentro da região demarcada dos vinhos verdes.

O *dataset* disponibilisado pela CVRVV contem os resultados físico-químicos e sensoriais, por boletim de análise, de todos os produtos que originam a atribuição dos seus Selos de Garantia, desde 2004 até 31 de dezembro de 2015, sendo que essa informação não contém qualquer tipo de elemento que permita a identificação do operador económico ou marca respetiva.

O *dataset* original era composto por 23.223 amostras de vinho verde branco, tinto ou rosado, assim como, de espumantes, vinagres, vinhos licorosos, aguardentes e vinho regional. A nossa análise recaiu sobre os vinhos verdes brancos com sub-região identificada, composto por 4.941 amostras de vinho verde branco, contendo 4 atributos com as caraterísticas gerais do produto, 13 atributos identificando as características físico-químicas e 8 atributos identificando as características organolépticas. Uma abordagem descritiva destes atributos pode ser encontrada no Anexo A1.

Os dados foram fornecidos pela CVRVV e apresentavam-se compactados num ficheiro Excel com 672.565 linhas, sendo que em média, cada amostra foi distribuída por 30 linhas. As características de cada variável, nomeadamente, nome, descrição, codificação e tipo de variável, apresentam-se descriminadas no Anexo A2.

4.1.1. Tratamento dos dados

O ficheiro rececionado em formato Excel, com 672.565 linhas foi reajustado, recorrendo ao *software* R, passando a dispor de 23.223 linhas, com uma amostra por linha, com 95 variáveis das quais se seleccionaram as 24 variáveis mais pertinentes e que tivessem menos de 1% de *missing values*. A pertinência da sua selecção justifica-se por ou não se aplicarem a vinhos verdes brancos com sub-região identificada ou terem uma representatividade inferior a 99%

dos dados. Depois desta alteração, as análises foram efectuadas no Knime, aplicando a técnica de *clustering*.

Foi efectuada a análise estatística por forma a obter as medidas de tendência central (média, mediana) e de dispersão, retirados através do *software* SPSS, podendo ser observados na seguinte Tabela 1. Para além desta análise verificamos também a existência de valores ausentes/omissos sendo pouco significativo nos atributos de maior relevância que se vai analisar.

	Válido	Ausente	Média	Mediana	Desvio Padrão	Mínimo	Máximo	Intervalo
AcidFix	4941	0	6.18	6.10	0.89	3.50	13.90	10.40
AcidTot	4941	0	6.55	6.50	0.88	3.80	14.20	10.40
AcidVolat	4941	0	0.30	0.29	0.10	0.05	1.08	1.03
AcidCitric	4940	1	0.31	0.29	0.11	-0.01	2.07	2.08
Cloret	4941	0	0.04	0.03	0.03	0.01	0.50	0.49
DioxEnxLiv	4941	0	30.16	29.00	14.01	0.00	339.00	339.00
DioxEnxTot	4941	0	115.75	112.00	31.69	9.00	481.00	472.00
ExtrNRed	4941	0	18.89	18.70	1.98	14.40	34.70	20.30
ExtrSecTot	4941	0	23.69	22.80	5.65	15.40	162.00	146.60
MassVol	4941	0	0.99	0.99	0.00	0.99	1.04	0.06
pH	4941	0	3.21	3.22	0.15	2.58	3.82	1.24
Sulfat	4940	1	0.48	0.46	0.15	0.17	1.93	1.76
TitAlcVolAdq	4940	1	11.90	12.00	0.95	8.60	14.90	6.30

Tabela 1 - Medidas da tendência central e de dispersão das características físico-químicas

O ficheiro de dados foi importado para o Knime através do nó *CSV Reader*. Posteriormente, e dado que os valores numéricos de diversas variáveis se encontram referidas em escalas diferentes, foi feita a normalização de dados, utilizando o nó *Normalizer* do Knime, transformando assim um valor numérico noutro valor numérico. Neste caso concreto, utilizou-se uma normalização gaussiana (normalização por padronização), que define um valor central e um valor de dispersão comuns para todas as variáveis, pois esta lida melhor com *outliers* (Gama; 2012). Esta normalização foi usada para todas as características físico-químicas, mas não para as características sensoriais. Os valores do segundo grupo estão na escala 1 a 10 com a interpretação pre-definida (ex. 6 e 7 representa “bom” etc. como se pode ver no Anexo A1).

4.2. Metodologia utilizada no estudo da semelhança entre vinhos

4.2.1. Método de agrupamento (*clustering*)

Para decidir se o vinho de um certo tipo e sub-região se assemelha a outro doutra sub-região usamos a técnica de agrupamento (*clustering*) e, em particular, o algoritmo *k-means*. A solução foi desenvolvida com o auxílio da plataforma Knime. Esta técnica foi explicada com maior detalhe no capítulo 3.1.

4.2.2. Como determinar o número de grupos

Como o algoritmo de *clustering* usado (*k-means*) não permite determinar o número de *clusters* ideal, usamos um procedimento iterativo que envolve o algoritmo EM, *Expectation Maximization*, segundo (Knime, 2015). Este procedimento começa com o mínimo número de *clusters* (1) e tenta-se aumentar esse número sempre que a medida de verossimilhança aumentar. De acordo com este procedimento o melhor número de *clusters* nos dados usados é 4. Assim, quando o *k-means* foi executado com o número de *clusters* = 4 para as 4.941 amostras (Tabela 2), foi obtido o seguinte resultado para os Dados I, com as sub-regiões conhecidas (ver Tabela 2).

Região	Nº Amostras	%			
Amarante	501	10.1			
Ave	471	9.5			
Baião	320	6.5			
Basto	391	7.9			
Cávado	431	8.7	Clusters	Nº Amostras	%
Lima	465	9.4	1	2076	42.02
Monção e Melgaço	1808	36.6	2	641	12.97
Paiva	155	3.1	3	1259	25.48
Sousa	399	8.1	4	965	19.53
Totais	4941	100	Totais	4941	100

Tabela 2 – Distribuição das amostras por região e Clusters gerados para Dados I

Em relação a Dados II, com as 14.355 amostras que incluem todos os dados, obtivemos:

			Clusters	Nº Amostras	%
Região	Nº Amostras	%	1	1208	8.4
Amarante	501	3.49	2	1876	13.1
Ave	471	3.28	3	2230	15.5
Baião	320	2.23	4	1574	11.0
Basto	391	2.72	5	1126	7.8
Cávado	431	3.00	6	1411	9.8
Lima	465	3.24	7	242	1.7
Monção e Melgaço	1808	12.59	8	330	2.3
Paiva	155	1.08	9	1802	12.6
Sousa	399	2.78	10	2250	15.7
S/região	9414	65.58	11	306	2.1
Totais	14355	100	Totais	14355	100

Tabela 3 – Distribuição das amostras por região e Clusters gerados para Dados II

4.2.3. Análise de Agrupamentos (*Clusters*) Gerados

Tentamos caracterizar da melhor maneira possível cada agrupamento de casos (*cluster*) gerado pelo programa, pois sem isso seria difícil pronunciarmo-nos sobre a utilidade dos mesmos. Procuramos obter uma caracterização que permitiria entender melhor quais as características que definem e distinguem cada *cluster*, em comparação com os outros.

O primeiro conceito que usamos nesta fase foi o conceito de ***centróide***. O centróide de um *cluster* tem a forma de um elemento, em que o valor de cada atributo é representado pela média (se se tratar de valores numéricos), ou moda se o atributo for categórico. Pode-se dizer que o centróide de um *cluster* representa esse *cluster*.

O facto que o centróide envolver todos os atributos pode não ser elucidativo, pois não chama a atenção dos atributos realmente importantes e discriminam bem este agrupamento em relação aos outros. Por esse motivo usamos outros meios complementares, como a caracterização de atributos usando o ***ganho da informação*** (*information gain*, IG).

Notamos que alguns atributos podem distinguir dois *clusters* diferentes como, por exemplo, *cluster* Ci e Cj. Outros, são úteis para distinguir o *cluster* Ci de todos os outros *clusters*. A primeira alternativa tem a desvantagem que para N *clusters*, o número de pares é N*(N-1)/2. Se o N for, por exemplo 11, o número de pares seria 55, o que é pouco prático. Para N=4, o número de pares já é aceitável (4*3/2=6), mas mesmo assim, optamos por uma outra alternativa, que nos parecia melhor. Nesta opção cada *cluster* é comparado com todos os outros. Deste modo, temos N problemas binários.

Assim, para calcular o valor de **ganho da informação** relativo aos atributos de *cluster* Ci, anotamos os casos desse *cluster* como casos positivos e todos os outros casos como negativos. O valor foi calculado usando o nó "*InformationGainCalculator*" do Knime. Para cada *cluster* Ci os atributos podem ser ordenados em ordem decrescente usando o valor desta medida. Nos resultados apresentados na secção seguinte limitamos a atenção aos atributos com *InfoGain*>0.1.

Podemos assim ver que, por exemplo, no *Cluster* 1 (ver Tabela 4) o atributo mais informativo é *SaborQualid*, pois tem o valor de *InfoGain*=0.649. O valor desse atributo no centróide é 7.133 pode ser considerado elevado face ao centróide dos outros *clusters* (5.964) e face à escala apresentada na Figura 11 e Figura 12 do Anexo A1.

Para obter a resposta a esta questão, usamos o *classificador* de tipo árvore de decisão (*Decision Tree Learner*) e depois transformamos a árvore em regras com *Decision Tree to Ruleset*. A seguir analisamos a regra com a maior cobertura de casos no *cluster* em questão. Depois disso, procuramos, para cada atributo informativo, a condição na regra que depois transcrevemos na tabela. Por exemplo, considerando um exemplo com *Cluster* 1 e atributo *AcidFix*, identificamos a seguinte condição na regra: *AcidFix* <= 0.75. Assim, os elementos deste *cluster* são caracterizados por valores *AcidFix* menores ou iguais a 0.75. Notamos ainda que o valor do centróide (-0.425) é bastante inferior a este limite e qual a sua localização face ao centróide dos outros grupos. Evidentemente, pela conjugação dos diversos atributos podemos encontrar outras caracterizações para este *cluster* (ver a secção seguinte.).

4.2.4. Descrição e Caracterização de *Clusters* gerados para Dados I

Nas 4 subsecções analisaremos cada agrupamento (*cluster*) separadamente. Cada secção inclui uma tabela que identifica os atributos mais informativos (coluna 2), o tipo de características (coluna 3), o valor de *InfoGain* (coluna 4), o centróide do *cluster* (coluna 5), o centróide dos outros *clusters* (coluna 6) e a condição gerada pelo sistema na regra com a maior cobertura (coluna 7). Os elementos nesta tabela estão ordenados pelo *InfoGain*.

Cluster 1 – Vinhos de qualidade superior e teor alcoólico elevado

Características: 2076 elementos, que representam 42.02 % dos dados totais.
A regra cobre 2046 amostras e estão correctamente cobertos 1975 amostras.

	Cluster 1	Características	InfoGain	Centróide	Centróide de Outros Clusters	Regra
1	SaborQualid	Sens.	0.649	7.133	5.964	
2	AromaQualid	Sens.	0.644	7.133	5.967	
3	SaborTipic	Sens.	0.637	7.137	6.007	
4	AromaTipic	Sens.	0.634	7.136	6.009	
5	AcidFix	Fís-Quím	0.148	-0.425	-0.202	AcidFix <= 0.75
6	AcidTot	Fís-Quím	0.131	-0.407	-0.217	
7	TitAlcVolAdq	Fís-Quím	0.123	0.447	-0.352	TitAlcVolAdq > -2.06
8	MassVol	Fís-Quím	0.102	-0.342	0.254	MassVol <= 1.11
9	CASTA	Identificação	0.068			
10	REGIAO	Identificação	0.058			
15	TIPOI	Identificação	0.033			

Tabela 4 - Dados relativos ao cluster 1 (Dados I)

Este agrupamento parece estar caracterizado por:

Características Físico-químicas

- Valor baixo de ***Acidez Fixa***, *AcidFix* (InfoGain=0.148). A regra estabelece que Acidez Fixa <= 0.75 e o valor deste atributo no centróide é Acidez Fixa = -0.425 que, comparado com o centróide dos outros grupos fica ainda mais abaixo, -0.202.

- Valor elevado de ***Título Alcoométrico Volúmico Adquirido***, *TitAlcVolAdq* (*InfoGain*=0.123). A regra estipula que *TitAlcVolAdq* > - 2.06, o valor deste atributo no centróide é TitAlcVolAdq=0.447.e é bastante superior face ao centróide dos outros grupos, -0.352.

- Valor baixo de ***Massa Volúmica***, *MassVol* (*InfoGain*=0.102). A regra estipula que *MassVol* <= 1.11 e o valor deste atributo no centróide *MassVol*= -0.423 é bastante inferior ao valor exigido, mesmo quando comparado com os centróides dos outros *clusters*, 0.254.

Características Sensoriais

- Possui informação pertinente no atributo ***Sabor Qualidade***, *SaborQualid* (*InfoGain*=0.649). Apesar de não estar definida nenhuma regra, no centróide possui uma nota de *SaborQualid*= 7.133, considerada boa pela escala apresentada na Figura 11 do Anexo A1, contra uma nota no centróide dos outros grupos de 5.964, considerada suficiente pela mesma escala.

- Na ***Aroma Qualidade***, tem também informação pertinente, *AromaQualid* (*InfoGain*=0.644). Apesar de não estar definida nenhuma regra, no centróide possui uma nota de *AromaQualid*= 7.133, considerada boa pela escala apresentada na Figura 11 do Anexo A1, contra uma nota no centróide dos outros grupos de 5.967, considerada suficiente pela mesma escala.

- Ao nível do ***Sabor Tipicidade*** possui informação pertinente, *SaborTipic* (*InfoGain*=0.637). Apesar de não estar definida nenhuma regra, no centróide possui uma nota de *SaborTipicid*= 7.137, considerada boa pela escala apresentada na Figura 12 do Anexo A1, contra uma nota no centróide dos outros grupos de 6.007, considerada também boa pela mesma escala.

- Na ***Aroma Tipicidade***, que classifica a sua tipicidade como vinho verde, a *AromaTipic* possui também bastante informação (*InfoGain*=0.634). A sua nota no centróide é de *AromaTipicid*= 7.136 que, segundo a escala apresentada na Figura 12 do Anexo A1, é considerada suficiente contra uma nota superior no centróide dos outros grupos 6.009, considerada boa pela mesma escala.

Observação (hipótese):

Ao nível das características físico-químicas possui ***Massa Volúmica*** (densidade) relativamente baixa face aos centróides dos outros *clusters*, indicando a presença de menos sólidos ou materiais insolúveis. Inversamente relacionado está o ***Título Alcoométrico Volúmico Adquirido*** (teor de alcool), elevado para vinho verde (12.33°), pois para a RDVV o valor mínimo é de 8°.

Quanto às características sensoriais deste *cluster* os atributos que classificam a ***Tipicidade***, tanto no Sabor como no Aroma, sinónimo identificador de um típico vinho verde, têm notas "boa"= 7, conforme escala (0 a 10) de atribuição representada na Figura 12 do Anexo A1. No respeitante à ***Qualidade***, tanto no Sabor como no Aroma, têm notas "boa"= 7, de acordo com a escala (0 a 10) de atribuição representada na Figura 11 do Anexo A1, bastante acima da

nota registada pela média dos outros *clusters*, 5.967. Este *cluster* é constituído maioritariamente por vinhos de Monção e Melgaço (52%), com maior representação face aos dados não agrupados (36.6%). A tabela abaixo indicada mostra a distribuição das amostras por regiões, podendo ser definido como incluindo vinhos de boa qualidade.

Sub-regiões	Nº Amostras	Amostras %	cluster 1	c1 %
Amarante	501	10.1	167	8.04
Ave	471	9.5	123	5.92
Baiao	320	6.5	84	4.05
Basto	391	7.9	143	6.89
Cavado	431	8.7	142	6.84
Lima	465	9.4	167	8.04
Monção e Melgaço	1808	36.6	1078	**51.93**
Paiva	155	3.1	37	1.78
Sousa	399	8.1	135	6.50
Totais	4941	100	2076	100

As cores utilizadas nestas tabelas para calcular o rácio são as seguintes:

Azul	Claro	< 0.8
	Normal	< 10%
Vermelho	Rosado	> 1.2
	Normal	> 10%

Cluster 2 – Vinhos com alguma acidez cítrica e de boa qualidade

Características: 641 elementos, que representam 12.97 % dos dados totais.
A regra cobre 162 amostras e estão correctamente cobertos 141 amostras.

	Cluster 2	Características	InfoGain	Centróide	Centróide de Outros Clusters	Regra
1	MassVol	Fís-Quím	0.460	1.322	-0.197	1.265 < MassVol <= 1.979
2	DioxEnxTot	Fís-Quím	0.307	1.213	-0.181	-0.324 < DioxEnxTot <= 2.706
3	ExtrSecTot	Fís-Quím	0.306	1.035	-0.154	ExtrSecTot <= 2.071
4	TitAlcVolAdq	Fís-Quím	0.281	-1.119	0.167	TitAlcVolAdq > -3.069
5	ExtrNRed	Fís-Quím	0.185	0.908	-0.135	
6	Cloret	Fís-Quím	0.140	0.626	-0.093	-0.279 < Cloret <= 5.226
7	SaborTipic	Sens.	0.100	6.153	6.594	
8	SaborQualid	Sens.	0.099	6.119	6.569	
9	AromaTipic	Sens.	0.099	6.153	6.595	
10	AromaQualid	Sens.	0.097	6.123	6.570	
11	AcidTot	Fís-Quím	0.087	0.384	-0.057	
12	AcidFix	Fís-Quím	0.081	0.409	-0.061	
13	DioxEnxLiv	Fís-Quím	0.073	0.587	-0.087	DioxEnxLiv > -1.332
14	Sulfat	Fís-Quím	0.060	0.531	-0.079	-1.379 < Sulfat <= 4.575
15	CASTA	Identificação	0.057			
16	AcidCitric	Fís-Quím	0.050	0.615	-0.092	AcidCitric > -2.036
17	REGIAO	Identificação	0.046			
18	TIPOI	Identificação	0.024			

Tabela 5 - Dados relativos ao cluster 2 (DadosI)

Este agrupamento parece estar caracterizado por:

Características Físico-químicas

- Valor elevado de ***Massa Volúmica***, *MassVol* (*InfoGain*=0.225). A regra estipula que este atributo varia entre 1.265 e 1.979 e o valor deste atributo no centróide *MassVol*= 1.322 é bastante superior ao valor exigido, sobretudo quando comparado com o centróide dos outros *clusters*, -0.197. O seu valor médio é de 0.990, ou seja, um pouco abaixo de 1.

- Valor bastante elevado de ***Dióxido de Enxofre Total***, *DioxEnxTot* (*InfoGain*=0.155). A regra estabelece que varie entre -3.24 e 2.706, sendo o seu valor deste atributo no centróide de *DioxEnxTot* = 1.213, logo, muito superior quando comparado com o centróide dos outros grupos, -0.181 .

- Valor alto de ***Extrato Seco Total***, *ExtrSecTot* (*InfoGain*=0.150). A regra estabelece que *ExtrSecTot* <= 2.071, o valor deste atributo no centróide é *ExtrSecTot* = 1.035, muito superior quando comparado com o centróide dos outros grupos, -0.154.

- Valor baixo de ***Título Alcoométrico Volúmico Adquirido***, *TitAlcVolAdq* (*InfoGain*=0.134). A regra estipula que *TitAlcVolAdq* > - 3.069, o valor deste atributo no centróide é *TitAlcVolAdq*=-1.118 e é bastante inferior face ao centróide dos outros grupos, 0.167.

- Valor elevado de ***Cloretos***, *Cloret* (*InfoGain*=0.140). A regra estabelece o valor de *Cloretos* = -0.279 < *Cloret* <= 5.226. O valor deste atributo no centróide (0.626) é elevado, relativamente ao valor apresentado pelo centróide dos outros grupos (-0.903).

- Valor bastante elevado de ***Dióxido de Enxofre Livre***, *DioxEnxLiv* (*InfoGain*=0.073). A regra estipula que seja superior a -1.332, sendo o valor deste atributo no centróide de *DioxEnxLiv* = 0.587 e muito superior quando comparado com o centróide dos outros grupos, -0.087.

- Valor elevado de ***Sulfatos***, *Sulfat* (*InfoGain*=0.060). A regra estabelece o valor de *Sulfatos* = -1.379 < *Sulfat* <= 4.575. O valor deste atributo no centróide (0.531) é elevado, relativamente ao valor apresentado pelo centróide dos outros grupos (-0.079).

- Valor elevado de ***Ácido Cítrico***, *AcidCitric* (*InfoGain*=0.050). A regra estabelece que o *Ácido Cítrico* > - 2.036. O valor deste atributo no centróide é *AcidCitric* = 0.615 que, comparado com o centróide dos outros grupos fica bastante acima, -0.092.

Características Sensoriais

- Ao nível do ***Sabor Tipicidade*** possui também informação pertinente, *SaborTipic* (*InfoGain*=0.100). Apesar de não estar definida nenhuma regra, no centróide possui uma nota de *SaborTipicid*= 6.153, um pouco inferior à nota do centróide dos outros grupos = 6.594, mas ambas consideradas como "boa" pela escala apresentada na Figura 12 do Anexo A1.

- Possui informação pertinente no atributo ***Sabor Qualidade***, *SaborQualid* (*InfoGain*=0.100). Apesar de não estar definida nenhuma regra, no centróide possui uma nota de *SaborQualid*= 6.153, um pouco inferior à nota do centróide dos outros grupos = 6.594, mas ambas consideradas como "boa" pela escala apresentada na Figura 11 do Anexo A1.

- Na ***Aroma Tipicidade***, que classifica a sua tipicidade como vinho verde, a *AromaTipic* possui também bastante informação (*InfoGain*=0.099). A sua nota no centróide é de *AromaTipicid*=

6.153, um pouco inferior à nota do centróide dos outros grupos = 6.595, mas ambas consideradas como "boa" pela escala apresentada na Figura 12 do Anexo A1.

- Na ***Aroma Qualidade***, tem também informação pertinente, *AromaQualid* (*InfoGain*=0.097). Apesar de não estar definida nenhuma regra, no centróide possui uma nota de *AromaQualid*= 6.123, um pouco inferior à nota do centróide dos outros grupos = 6.570, mas ambas consideradas como "boa" pela escala apresentada na Figura 11 do Anexo A1.

Observação (hipótese):

A elevada massa volúmica contrapõe com o baixo teor alcoólico. Denota possuir também alguma acidez cítrica que, provavelmente, poderá conferir alguma tipicidade no aroma e no sabor. Apresenta boa qualidade e tipicidade no seu centróide, conforme escala de valores da Figura 11 e Figura 12 do Anexo A1. O *cluster* é constituído maioritariamente por vinhos de 4 sub-regiões (64%), Amarante, Ave, Sousa e Cávado. A sub-região Monção e Melgaço desceu face aos dados não agrupados (36.6%) como pode ser constatado pela seguinte tabela:

Sub-regiões	Nº Amostras	Amostras %	cluster 2	c2%
Amarante	501	10.1	99	15.44
Ave	471	9.5	98	15.29
Baiao	320	6.5	42	6.55
Basto	391	7.9	51	7.96
Cavado	431	8.7	93	14.51
Lima	465	9.4	56	8.74
Monção e Melgaço	1808	36.6	119	**18.56**
Paiva	155	3.1	11	1.72
Sousa	399	8.1	72	11.23
Totais	4941	100	641	100

Cluster 3 – Vinhos com elevada acidez

Características: 1259, que representam 25.48 % dos dados totais.
A regra cobre 1235 amostras e estão correctamente cobertas 1187 amostras.

	***Cluster* 3**	**Características**	**InfoGain**	**Centróide**	**Centróide de Outros Clusters**	**Regra**
1	SaborQualid	Sens.	0.463	5.886	6.725	SaborQualid <= 6.5
2	AromaQualid	Sens.	0.455	5.888	6.726	
3	SaborTipic	Sens.	0.422	5.933	6.744	
4	AromaTipic	Sens.	0.414	5.936	6.743	
5	AcidFix	Fís-Quím	0.175	-0.512	0.175	AcidFix <= 0.529
6	AcidTot	Fís-Quím	0.164	-0.523	0.179	
7	ExtrSecTot	Fís-Quím	0.085	-0.332	0.114	ExtrSecTot <= 1.267
8	ExtrNRed	Fís-Quím	0.081	-0.420	0.144	ExtrNRed <= 2.309
9	MassVol	Fís-Quím	0.051	-0.289	0.099	MassVol <= 1.631
10	TIPOI	Identificação	0.030			
11	CASTA	Identificação	0.026			
17	REGIAO	Identificação	0.007			

Tabela 6 - Dados relativos ao cluster 3 (Dados I)

Este agrupamento parece estar caracterizado por:

Características Físico-químicas

- Valor baixo de ***Acidez Fixa***, *AcidFix* (*InfoGain*=0.175). A regra estabelece que Acidez Fixa <= 0.529 e o valor deste atributo no centróide é Acidez Fixa = -0.512 que, comparado com o centróide dos outros grupos fica ainda mais abaixo, 0.175.
- Valor alto de ***Extrato Seco Total***, *ExtrSecTot* (*InfoGain*=0.085). A regra estabelece que ExtrSecTot <= 1.267, o valor deste atributo no centróide é ExtrSecTot = -0.332, inferior quando comparado com o centróide dos outros grupos, 0.114.
- Valor baixo de ***Extrato Não Redutor***, *ExtrNRed* (*InfoGain*=0.081). A regra estabelece que *ExtrNRed* <= 2.309, o valor deste atributo no *centróide* é ExtrNRed =-0.420 que, comparado com o centróide dos outros grupos fica ainda mais abaixo, 0.144.
- Valor de ***Massa Volúmica***, *MassVol* (*InfoGain*=0.051). A regra estipula que a Massa Volúmica deverá ser *MassVol* <= 1.631, sendo o valor deste atributo no centróide de *MassVol* =-0.289, abaixo do registado no centróide dos outros *clusters*, *MassVol* =0.099.

Características Sensoriais

- Valor reduzido de ***Sabor Qualidade***, *SaborQualid* (*InfoGain*=0.463). A regra estipula que *SaborQualid* <= 6.5, o valor deste atributo no centróide é *SaborQualid* =5. 886, nota suficiente segundo escala da Figura 11 do Anexo A1, inferior à nota do centróide dos outros grupos, 6.725, definida como boa pela mesma escala.
- Valor reduzido de ***Aroma Qualidade***, *AromaQualid* (InfoGain=0.455). Apesar de não estar definida nenhuma regra, o valor deste atributo no centróide é *AromaQualid* =5.888, nota suficiente segundo escala da Figura 11 do Anexo A1, inferior à nota do centróide dos outros grupos, 6.726, definida como boa pela mesma escala.
- Valor reduzido de ***Sabor Tipicidade***, *SaborTipic* (*InfoGain*=0.422). Apesar de não estar definida nenhuma regra, o valor deste atributo no centróide é SaborTipic =5.933, nota suficiente segundo escala da Figura 12 do Anexo A1, inferior à nota do centróide dos outros grupos, 6.744, definida como boa pela mesma escala.
- Valor reduzido de ***Aroma Tipicidade***, *AromaTipic* (*InfoGain*=0.414). Apesar de não estar definida nenhuma regra, o valor deste atributo no centróide é *AromaTipic* =5.936, nota suficiente segundo escala da Figura 12 do Anexo A1, inferior à nota do centróide dos outros grupos, 6.743, definida como boa pela mesma escala.

Observação (hipótese):

Os valores da Acidez Fixa são bastante elevados pois a sua média é de 5.72 e o valor mínimo permitido pelos estatutos da Região Demarcada dos Vinhos Verdes (Art. 11º) é de 5.4. A corroborar esta característica temos as notas atribuídas nos item que definem a sua tipicidade e qualidade com nota 5 na classificação da tipicidade (valor mínimo a partir do qual se considera

ter alguma tipicidade) e, no caso da qualidade, seja de aroma seja de sabor, temos nota "suficiente", segundo a escala da Figura 11 do Anexo A1. Este *cluster* é constituído maioritariamente por vinhos da sub-região de Monção e Melgaço (34%), precedido, de longe, pela sub-região do Ave (12.39%), que aumentou face aos dados não agrupados (9.5%) como pode ser constatado pela seguinte tabela:

Sub-regiões	Nº Amostras	Amostras %	cluster 3	c2%
Amarante	501	10.1	123	9.77
Ave	471	9.5	156	12.39
Baiao	320	6.5	96	7.63
Basto	391	7.9	98	7.78
Cavado	431	8.7	98	7.78
Lima	465	9.4	125	9.93
Monção e Melgaço	1808	36.6	425	33.76
Paiva	155	3.1	50	3.97
Sousa	399	8.1	88	6.99
Totais	4941	100	1259	100

Cluster 4 – Vinhos com Densidade e Extrato Seco Total elevados

Características: 965 elementos, que representa 19.5% dos dados totais.
A regra cobre 1243 amostras e estão correctamente cobertos 1202 amostras.

	***Cluster* 4**	**Características**	**InfoGain**	**Centróide**	**Centróide de Outros Clusters**	**Regra**
1	MassVol	Fís-Quím	0.225	1.322	-0.057	1.265 < AND MassVol <= 1.979
2	DioxEnxTot	Fís-Quím	0.155	1.213	0.087	-0.324 < DioxEnxTot <= 2.706
3	ExtrSecTot	Fís-Quím	0.150	1.035	-0.037	ExtrSecTot <= 2.071
4	TitAlcVolAdq	Fís-Quím	0.134	-1.118	0.065	TitAlcVolAdq > -3.069
5	ExtrNRed	Fís-Quím	0.085	0.908	-0.126	
6	Cloret	Fís-Quím	0.072	-0.093	0.012	
16	CASTA	Identificação	0.023			
17	REGIAO	Identificação	0.023			
19	TIPOI	Identificação	0.009			

Tabela 7 - Dados relativos ao cluster 4 (DadosI)

Este agrupamento parece estar caracterizado por:

Características Físico-químicas

- Valor elevado de ***Massa Volúmica***, *MassVol* (*InfoGain*=0.225). A regra estipula o seu valor variando entre 1.265 < *MassVol* <= 1.979. O seu valor no centróide é *MassVol* =1.322 enquanto que o do centróide dos outros grupos é *MassVol* = -0.057.
- Valor elevado de ***Dióxido de Enxofre Total***, *DioxEnxTot* (InfoGain=0.155). A regra estipula que -0.324 < *DioxEnxTot* <= 2.706. O valor deste atributo no centróide é *DioxEnxTot* =1.213 e o do centróide dos outros grupos é *DioxEnxTot* = 0.087.

- Apresenta também valor elevado de ***Extrato Seco Total***, *ExtrSecTot* (*InfoGain*=0.150). A regra estipula que *ExtrSecTot* <= 2.071. O seu valor no centróide é *ExtrSecTot* = 1.035 enquanto o do centróide dos outros grupos é *ExtrSecTot* = -0.037.
- Valor bastante reduzido de ***Título Alcoométrico Volúmico Adquirido***, *TitAlcVolAdq* (*InfoGain*=0.290). A regra estipula que *TitAlcVolAdq* <= 1.051. O seu valor no centróide é *TitAlcVolAdq* = -1.118 enquanto o do centróide dos outros grupos é TitAlcVolAdq = 0.065.

Características Sensoriais

Não apresenta valores sensoriais a reportar.

Observação (hipótese):

Possui Massa Volúmica muito elevada (1.322), muito acima do valor de referência da água, que é 1. Este atributo está inversamente relacionado com o baixo teor alcoólico (Título Alcoométrico Volúmico Adquirido). Para além disso, são identificados vinhos com um elevado Extrato Seco Total, tornado-se mais espesso na boca. Em resumo, o *cluster* é caracterisado por vinhos com nota de qualidade suficiente, de acordo com a escala da Figura 11 do Anexo A1, e com nota 6 na escala da tipicidade, na Figura 12 do Anexo A1. Este *cluster* reune a maior parte dos vinhos das diversas sub-regiões pois apesar de o vinho com maior peso ser Monção e Melgaço, com 19.27%, que diminuiu face aos dados não agrupados (36.6%), as sub-regiões Baião, Basto, Lima, Paiva e Sousa aumentaram a sua presença face aos dados não agrupados como pode ser constatado pela seguinte tabela:

Sub-regiões	Nº Amostras	Amostras %	c4	c4%
Amarante	501	10.1	112	11.61
Ave	471	9.5	94	9.74
Baiao	320	6.5	98	10.16
Basto	391	7.9	99	10.26
Cavado	431	8.7	98	10.16
Lima	465	9.4	117	12.12
Monção e Melgaço	1808	36.6	186	**19.27**
Paiva	155	3.1	57	5.91
Sousa	399	8.1	104	10.78
Totais	4941	100	965	100

4.2.5. Conclusão do estudo com dados reduzidos (Dados I)

Este estudo permitiu avançar na resposta a algumas das questões levantadas inicialmente, nomeadamente, se os vinhos de uma determinada sub-região se assemelham a vinhos de outra sub-região.

Para isso foram identificados *clusters* com elementos possuindo características semelhantes. Apesar da dimensão do universo dos vinhos estudado ser relativamente pequena permitiu

identificar quatro *clusters* com características bem identificadas e referindo-se tanto a sub-regiões específicas como ao conjunto de quase todas as regiões, abaixo identificadas:

- ***Cluster* 1 -** Este *cluster* é constituído maioritariamente por vinhos com ***baixa acidez e teor alcoólico elevado***, com maior representação da sub-região de ***Monção*** e ***Melgaço*** (52%) face aos dados não agrupados (36.6%).
- ***Cluster* 2 -** O *cluster* é constituído maioritariamente por vinhos com ***acidez elevada e boa qualidade*** de 4 sub-regiões (40%), ***Amarante***, ***Ave, Cávado e Sousa***. Embora ***Monção*** e ***Melgaço*** tenha uma representação significativa (18.56%) desceu face aos dados não agrupados (36.6%).
- ***Cluster* 3** – A característica que mais sobressai é a ***elevada acidez***, influenciando negativamente as classificações atribuídas tanto quanto à *Tipicidade* como quanto à *Qualidade*.
- ***Cluster* 4** – Vinhos com ***elevada densidade***, influenciada pelos níveis de açúcar, e relativo **baixo teor alcoólico**, identificador de um conjunto de vinhos com características comuns a oito das nove sub-regiões. Embora ***Monção*** e ***Melgaço*** tenha uma representação significativa (19.27%) desceu face aos dados não agrupados (36.6%).

Esperamos que, com esta identificação dos *clusters* e sua caracterização, permita aos técnicos da especialidade avançar com a análise mais detalhada para ver se foram identificados alguns grupos de interesse e, desse modo, melhorar, pelo menos em alguns aspetos, o conhecimento da área.

4.2.6. Descrição e Caracterização de *Clusters* gerados para Dados II

Nas 11 subsecções analisaremos cada agrupamento (*cluster*) separadamente. Cada secção inclui uma tabela que mostra os atributos mais informativos (coluna 2), o tipo de características (coluna 3), o valor de *InfoGain* (coluna 4). Os elementos nesta tabela estão ordenados pelo *InfoGain*.

Na tabela, para além do centróide existente na coluna 5 e do centróide de outros *clusters* (coluna 6), inclui ainda a condição gerada pelo sistema na regra com a maior cobertura (coluna 7).

Como a descrição dos onze *clusters* é muito extensa para o presente capítulo iremos aqui apresentar somente dois dos *clusters*, aqueles cujas sub-regiões têm maior representatividade, o *cluster* 9 e o *cluster* 10, sendo os remanescentes *clusters* remetidos para o Anexo A3.

Cluster 9 - Vinhos com Densidade baixa

Características: 1802 elementos, que representam 12.6% de dados.
A regra cobre 1737 amostras e estão correctamente cobertos 1633 amostras.

	Cluster 9	Características	InfoGain	Centróide	Centróide de Outros Clusters	Regra
1	MassVol	Fis.-Quim.	0.333	-0.924	0.194	MassVol <= -0.024
2	SaborQualid	Sens.	0.291	5.956	6.105	SaborQualid <= 6.5
3	AromaQualid	Sens.	0.286	5.957	6.107	
4	TitAlcVolAdq	Fis.-Quim.	0.276	0.865	-0.303	TitAlcVolAdq > -0.39
5	SaborTipic	Sens.	0.257	5.977	6.169	SaborTipic > 5.5
6	AromaTipic	Sens.	0.255	5.977	6.170	
7	TitAlcVolTot	Fis.-Quim.	0.208	0.722	-0.289	
8	ExtrSecTot	Fis.-Quim.	0.186	-0.697	0.062	
9	RelAlcPeso_ExtNRed	Fis.-Quim.	0.156	0.473	-0.179	RelAlcPeso_ExtNRed > -0.48
10	DioxEnxTot	Fis.-Quim.	0.148	-0.712	0.092	DioxEnxTot <= 1.89
11	Cloret	Fis.-Quim.	0.114	-0.293	0.536	Cloret <= 6.56
12	AcucarRed	Fis.-Quim.	0.074	-0.362	-0.006	
13	DioxEnxLiv	Fis.-Quim.	0.067	-0.443	0.021	DioxEnxLiv <= 3.12
14	REGIAO	Identificação	0.064			
15	ExtrNRed	Fis.-Quim.	0.064	-0.409	0.108	
16	AcucarTot	Fis.-Quim.	0.063	-0.206	0.118	
22	TIPOI	Identificação	0.016			
23	CASTA	Identificação	0.013			

Tabela 8 - Dados relativos ao cluster 9 (Dados II)

Este agrupamento parece estar caracterizado por:

- Valor de ***Massa Volúmica*** baixo, *MassVol* (*InfoGain*=0.333). A regra estipula que *MassVol* <= -0.024 e o valor deste atributo no centróide é MassVol =-0.924.
- Valor baixo de ***Sabor Qualidade***, *SaborQualid* (*InfoGain*=0.291). A regra estipula que *SaborQualid* <= 6.5 e o valor deste atributo no centróide é *SaborQualid*=5.956 quando o valor no centróide dos outros *clusters* é de 6.105.
- Valor elevado de ***Título Alcoométrico Volúmico Adquirido***, *TitAlcVolTot* (*InfoGain*=0.276). A regra estipula que *TitAlcVolAdq* > -0.39 e o seu valor no centróide é *TitAlcVolAdq*=0.865.
- Valor baixo de ***Sabor Tipicidade***, *SaborTipic* (*InfoGain*=0.257). A regra estipula que *SaborTipic* > 5.5 e o seu valor no centróide é significativamente baixo, *SaborTipic* =5.977.
- Valor elevado de ***Relação Alcool/Peso e Extrato Não Redutor***, *RelAlcPeso_ExtNRed* (*InfoGain*=0.156). A regra estipula que *RelAlcPeso_ExtNRed* > -0.48 e o seu valor no centróide é *RelAlcPeso_ExtNRed* =0.473.
- Valor baixo de ***Dióxido de Enxofre Total***, *DioxEnxTot* (*InfoGain*=0.148). A regra estipula que *DioxEnxTot* <=1.89 e o valor deste atributo no centróide é *DioxEnxTot* =-0.712.
- Valor baixo de ***Cloretos***, *Cloret* (*InfoGain*=0.114). A regra estipula que *Cloret* <= 6.56 e o valor deste atributo no centróide é *Cloret* =-0.293.

Observação (hipótese):

Cluster cujos valores apresentam ganhos e presença significativa ao nível da Massa Volúmica baixa, sinónimo da presença de elevado teor alcoólico, confirmado pelo atributo Título Alcoométrico Volúmico Adquirido. A média deste último atributo é 11.99°, grau bastante elevado para vinhos verdes. Talvez por isso, tenham atribuído aos vinhos deste *cluster* uma nota média baixa, 5.96 (Suficiente) nos itens Sabor Qualidade e Sabor Tipicidade. Deste modo, estes vinhos devem ser classificados como vinhos de qualidade inferior.

Este *cluster* é constituído por vinhos sem sub-região identificada (41%), que desceram face aos dados não agrupados (65.58%) e por um aumento da presença de todas as sub-regiões, também comparativamente com os dados não agrupados, conforme pode ser constatado pela distribuição das amostras por regiões na seguinte tabela:

Sub-regiões	Nº Amostras	Amostras %	c9	c9%
Amarante	501	3.49	97	5.38
Ave	471	3.28	88	4.88
Baiao	320	2.23	102	5.66
Basto	391	2.72	80	4.44
Cavado	431	3.00	69	3.83
Lima	465	3.24	88	4.88
Monção e Melgaço	1808	12.59	427	**23.70**
Paiva	155	1.08	46	2.55
Sousa	399	2.78	69	3.83
S/região	9414	65.58	736	40.84
Totais	14355	100	1802	100

Cluster 10 – Vinhos com Sabor e Teor Alcoólico elevado

Características: 2250 elementos, que representam 15.7 % de dados.
A regra cobre 2180 amostras e estão correctamente cobertos 2097 amostras.

	Cluster 10	Características	InfoGain	Centróide	Centróide de Outros Clusters	Regra
1	SaborQualid	Sens.	0.643	7.105	5.897	
2	SaborTipic	Sens.	0.642	7.109	5.966	
3	AromaTipic	Sens.	0.642	7.109	5.967	AromaTipic > 6.5
4	AromaQualid	Sens.	0.635	7.104	5.899	
5	TitAlcVolAdq	Fis.-Quim.	0.402	1.154	-0.331	TitAlcVolAdq > -0.39
6	MassVol	Fis.-Quim.	0.323	-0.919	0.194	MassVol <= 0.92
7	TitAlcVolTot	Fis.-Quim.	0.303	0.939	-0.311	TitAlcVolTot > -1.12
8	AcidFix	Fis.-Quim.	0.216	-0.739	0.155	AcidFix <= 0.27
9	RelAlcPeso_ExtNRed	Fis.-Quim.	0.201	0.533	-0.185	RelAlcPeso_ExtNRed > -1.07
10	REGIAO	Identificação	0.177			
11	AcidTot	Fis.-Quim.	0.166	-0.677	0.159	
12	CASTA	Identificação	0.159			
13	Cloret	Fis.-Quim.	0.126	-0.307	0.538	Cloret <= 5.98
14	ExtrSecTot	Fis.-Quim.	0.118	-0.495	0.042	
15	DioxEnxTot	Fis.-Quim.	0.071	-0.353	0.056	
16	TitAlcVolTotReg	Fis.-Quim.	0.071	0.190	-0.043	
17	pH	Fis.-Quim.	0.0703	0.521	-0.139	
18	TIPOI	Identificação	0.0695			
19	AcucarTot	Fis.-Quim.	0.0647	-0.208	0.030	
20	AcidVolat	Fis.-Quim.	0.0591	0.442	0.030	
21	AcucarRed	Fis.-Quim.	0.0529	-0.188	-0.023	
22	ExtrNRed	Fis.-Quim.	0.0516	-0.338	0.101	ExtrNRed <= 1.88

Tabela 9 - Dados relativos ao cluster 10 (Dados II)

Este agrupamento parece estar caracterizado por:

- Valor elevado de ***Aroma Tipicidade***, *AromaTipic* (*InfoGain*=0.642). A regra estipula uma *AromaTipic* > 6.5 e o valor deste atributo no centróide é elevado, *AromaTipic* =7.109, comparativamente com o centróide dos outros clusters, 5.967.

- Valor elevado de ***Título Alcoométrico Volúmico Adquirido***, *TitAlcVolTot* (*InfoGain*=0.402). A regra estipula que *TitAlcVolAdq* > -0.39 e o valor deste atributo no centróide é *TitAlcVolAdq*=1.154.

- Valor de ***Massa Volúmica*** baixo, *MassVol* (*InfoGain*=0.323). A regra estipula que *MassVol* <= 0.92 e o valor deste atributo no centróide é *MassVol* =-0.919.

- Valor elevado de ***Título Alcoométrico Volúmico Total***, *TitAlcVolTot* (*InfoGain*=0.303). A regra estipula que *TitAlcVolTot* > -1.12 e o valor deste atributo no centróide é *TitAlcVolTot* = -0.939.

- Valor reduzido de ***Acidez Fixa***, *AcidFix* (*InfoGain*=0.216). A regra estipula que Acidez Fixa <= 0.27 e o valor deste atributo no centróide é Acidez Fixa =-0.739.

- Valor elevado de ***Relação Alcool/Peso e Extrato Não Redutor***, *RelAlcPeso_ExtNRed* (*InfoGain*=0.201). A regra estipula *RelAlcPeso_ExtNRed* > -1.07 e o valor deste atributo no centróide é *RelAlcPeso_ExtNRed* =0.533.

- Valor baixo de ***Cloretos***, *Cloret* (*InfoGain*=0.126). A regra estipula que *Cloret* <= 5.98 e o valor deste atributo no centróide é relativamente baixo, *Cloret* =-0.307.

Observação (hipótese):

Cluster com valores elevados nos atributos organolépticos, dos quais se destaca o da Aroma Tipicidade, considerado na escala de medida como Bom, com 7.11 pontos. Apresenta médias da Massa Volúmica, aparada a 5% ou não aparada, iguais à da água (1.0), valor indicativo de baixo teor alcoólico (Título Alcoométrico Volúmico Adquirido) e com baixa presença de Acidez Fixa e de Cloretos. Deste modo, este *cluster* é caracterizado por vinhos de qualidade superior.

Este *cluster* é constituído maioritariamente por vinhos de Monção e Melgaço (41%), que subiu significativamente face aos dados não agrupados. Note-se que os vinhos sem sub-região baixaram consideravelmente face aos dados não agrupados, como pode ser constatado pela distribuição das amostras por regiões na seguinte tabela:

Sub-regiões	Nº Amostras	Amostras %	c10	c10%
Amarante	501	3.49	140	6.22
Ave	471	3.28	80	3.56
Baiao	320	2.23	67	2.98
Basto	391	2.72	119	5.29
Cavado	431	3.00	67	2.98
Lima	465	3.24	99	4.40
Monção e Melgaço	1808	12.59	927	**41.20**
Paiva	155	1.08	27	1.20
Sousa	399	2.78	89	3.96
S/região	9414	65.58	635	**28.22**
Totais	14355	100	2250	100

4.2.7. Conclusão do estudo com dados completos (Dados II)

Este segundo estudo assemelha-se ao primeiro, descrito na secção 4.2.5., mas há uma diferença importante. Na secção acima referida usámos um subconjunto com a região conhecida. Para este conjunto de dados o método usado (descrito na secção 4.2.4.) indicou que o número ideal de *clusters* é 4. Os resultados deste estudo foram apresentados na secção anterior.

Nesta secção decidimos usar todos os dados, mesmo quando a região era desconhecida. Para esta segunda situação, o método indicou 11 *clusters* com características semelhantes.

Apesar de a maioria dos *clusters* reportados se referir a vinhos sem sub-região identificada, surgem alguns casos bem definidos caracterizados, um deles por vinhos com sabor e teor alcoólico elevado com presença maioritária da sub-região de Monção e Melgaço. Nota-se que um resultado semelhante foi obtido no estudo descrito na secção 4.2.4.

Infelizmente, em muitos dos casos a região não se encontra identificada nos dados, e assim, não podemos facilmente dar a resposta à questão colocada. No entanto, esperamos que, apesar disso, os *clusters* identificados, junto com a caracterização fornecida, permita aos técnicos da especialidade avançar com a análise mais detalhada para verificar se foram identificados alguns grupos de interesse e, desse modo, melhorar, pelo menos em alguns aspetos, o conhecimento da área.

5. Estudo sobre a avaliação de um painel de provadores de vinhos

O objetivo principal consiste em efectuar um estudo de caracterização dos provadores a partir duma base de dados de vinhos, com informação recolhida ao longo de oito anos (2007-2015), e a análise das características organolépticas de um painel de provadores constituídos por 15 enólogos. Estes especialistas classificam o vinho com um valor, numa escala de pontuação entre 0 e 20, tendo por base um vinho de referência.

5.1. Descrição dos dados e Pré-processamento

O *dataset* disponibilisado pela Sograpе contem os resultados físico-químicos e sensoriais, por amostra de prova, realizadas entre 2007 até 2015, sendo que essa informação não contém qualquer tipo de elemento que permita a identificação do operador económico ou marca respetiva.

O *dataset* original era composto por 1.272 amostras de prova de diversas regiões (Alentejo, Dão, Douro, Vinhos Verdes e Sem Denominação de Origem), branco, tinto ou rosado, assim como, vinhos licorosos (vinho do Porto) e vinhos regionais. A nossa análise recaiu sobre os ***vinhos brancos*** das regiões do Alentejo, Dão, Douro e Vinhos Verdes, com 336 amostras de prova, contendo 7 atributos a identificar as caraterísticas gerais do produto, 30 atributos referindo-se às características físico-químicas, 19 atributos relativos às características organolépticas e 14 atributos contendo informação estatística descritiva da amostra. A descrição das características físico-químicas pode ser encontrada no Anexo B1.

As características de cada variável, nomeadamente, *nome*, *descrição*, *codificação* e *tipo de variável*, apresentam-se discriminadas no Anexo B2.

Do ficheiro foi seleccionado um *dataset* composto por um conjunto de 336 provas de vinho em que participaram 15 provadores, atribuindo uma Nota que pode variar entre 0 e 20, sendo que 285 provas é o número máximo em que um único provador participou.

Verificou-se existir uma amostra defeituosa com o número 315, tendo recebido a Nota zero (0), atribuída por cinco provadores. Por ser uma observação com grande influência, perturbando demasiadamente os dados, foi retirada do universo a estudar. Os provadores P888 e P938 participam num número de provas inferior a 50, mas serão mantidos conforme se poderá observar na Tabela 10.

Após esta seleção foram ainda analisadas as características físico-químicas de modo a prevalecerem somente as variáveis com mais de metade dos valores preenchidos, ou seja, mais de 1134 registos com informação preenchida. As variáveis abaixo indicadas na Tabela 10 foram retiradas por não cumprirem esta condição:

Variáveis	registos sem	%
Intensidade	2163	95.4
Tonalidade	2228	98.2
GlucoseFrutose	2140	94.4
Taninos-gL	1648	72.7
Ca-mgL	2138	94.3
AcetatodeEtilo-mgL	2168	95.6
Sorbatodepotassio-mgL	2236	98.6

Tabela 10 - Variáveis sem informação para o estudo

5.2. Análise estatística das Notas dos provadores

5.2.1. Cálculo das medidas de tendência central (média) e de dispersão das Notas

Neste ponto pretende-se estudar o comportamento dos provadores de modo a identificar os que atribuem Notas acima (ou abaixo) da média.

As variáveis que constam na tabela abaixo identificam nas respectivas colunas:

1 - Provadores;
2 - nr: o número de participação em provas de análise de vinhos;
3 - Média: a média das Notas do provador;
4 - MédiaO: a média das Notas dos outros provadores;
5 - Média-MédiaO: a diferença entre a média da Nota do provador e a média das Notas dos outros provadores;
6 - Desvio Padrão: desvio padrão do provador;
7 - Desvio PadrãoO: desvio padrão dos outros provadores;
8 - *p-value* (*2-tailed*): *p-value* do provador.

Foi efectuada a análise estatística por forma a obter as medidas de tendência central (média) e de dispersão (desvio padrão), podendo os resultados ser observados nas primeiras 7 colunas da Tabela 11 da página seguinte.

Provador	nr	Média	MédiaO	Média-MédiaO	Desvio Padrão	Desvio PadrãoO	p-value
p888	25	14.68	14.058	0.622	1.737	1.006	0.139
p444	209	14.361	14.066	0.295	1.332	1.078	0
p467	239	14.285	14	0.285	1.886	1.402	0
p800	181	14.334	14.069	0.265	1.345	1.343	0.002
p555	93	14.366	14.142	0.224	1.606	1.169	0.073
p284	220	14.116	13.933	0.183	1.601	1.522	0.814
p441	285	14.126	14.017	0.109	1.952	1.304	0.825
p938	10	12.8	12.74	0.06	1.687	1.819	0.049
p384	83	13.928	13.887	0.041	1.949	1.435	0.801
p735	50	13.806	13.971	-0.165	2.757	1.716	0.174
p112	183	13.618	13.838	-0.22	2.094	1.392	0.147
p736	207	13.865	14.087	-0.222	2.332	1.4	0.009
p828	76	13.73	13.985	-0.255	1.718	1.174	0.128
p911	210	13.555	13.932	-0.377	1.739	1.423	0.148
p924	214	13.465	13.922	-0.457	2.051	1.349	0.004
Médias	**152.4**	**13.852**			**1.235**		

Tabela 11 - Medidas da tendência central e de dispersão dos provadores

Na Tabela 11 nota-se que alguns provadores têm tendência para atribuir a Nota acima ou abaixo da média dos outros. A questão surge se estas diferenciações são estatisticamente significativas. Esta questão motivou-nos para aplicar um teste estatístico, no qual comparamos as Notas de um dado provador com as médias dos outros provadores. Aplicou-se o teste *t de Student* (Marôco, 2014; Pestana, 2014). Foi efetuada uma amostra emparelhada de cada um dos provadores com a média dos outros provadores (ver anexo B3).

Deste modo, conclui-se que os provadores que atribuem Notas ***significativamente inferiores à média*** dos outros provadores são o ***P924*** e ***P736***, pois em ambos os casos os avaliadores atribuem Nota inferior à média e o *p-value* é menor que 0.01. Estes casos estão assinalados na Tabela 11, a cor azul. Nota-se, que muito embora os provadores ***P828*** e ***P911*** atribuem também Nota inferior à média, o *p-value* excede o patamar de 0.01 e, por esse motivo, estes casos não são considerados como significativamente inferiores à média.

Os provadores que atribuem Notas ***significativamente superiores à média*** são o ***P444***, o ***P467*** e o ***P800***, sendo que o primeiro regista as Notas com maior viés positivo em relação aos outros provadores, atribuindo as Notas mais elevadas. Em todos os três casos o *p-value* é menor que 0.01 e, por esse motivo, as diferenças são consideradas estatisticamente significativas. Estes casos estão assinalados na Tabela 11 a cor vermelha.

O provador ***P888*** parece ser um caso específico. O viés positivo dele é maior que os outros provadores referidos acima (i.e. P444, etc.), mas apesar disso o teste estatístico não assinalou a diferença como significativa (*p-value* é 0.139, i.e. acima de 0.01). No nosso entender isso deve-

se ao facto desse provador ter participado em poucas provas (26) comparativamente com a média dos outros provadores (152.4).

Os remanescentes provadores atribuem Notas comparáveis, i.e., não significativamente diferentes às da média dos outros provadores.

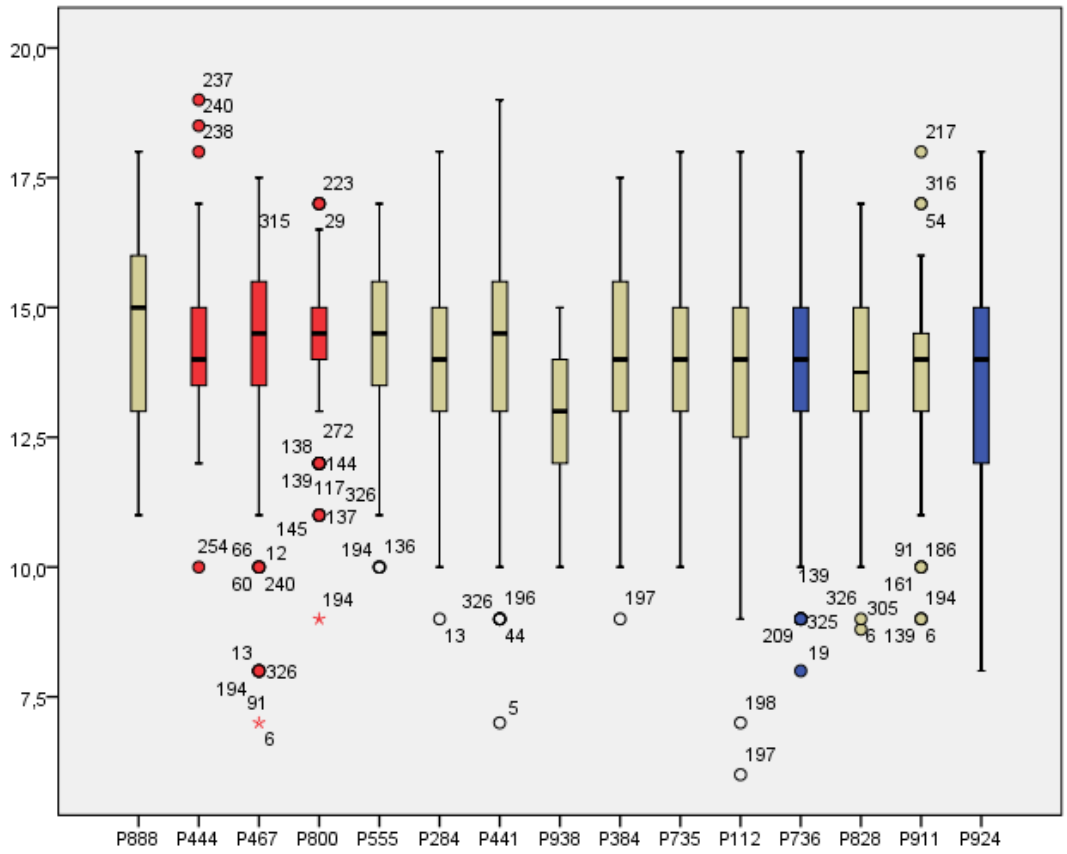

Figura 9 - Box and whisker plot dos provadores

De acordo com a Tabela 11 e a Figura 9 identificam-se os provadores que têm a tendência de atribuir Notas mais altas à esquerda (P888, P444 e P467) e os provadores mais que atribuem Notas mais baixas à direita (P924 e P911). Dos primeiros, o ***P888***, apesar de participar em apenas 26 provas, tem a média de Notas mais alta.

Podemos constatar a existência de diversos *outliers*, tanto moderados como severos, pelo que o estudo podia ser efetuado de novo, com uma eliminação de alguns dos *outliers* mais severos, de modo a não subverter a análise final dos dados. O ***P444*** apresenta 4 *outliers* severos, e dos provadores que atribuem as Notas mais baixas, o ***P911*** tem diversos *outliers* com Notas inferiores e também superiores à média. No Anexo B3 apresenta-se uma tabela a identificar as participações e ausências de cada provador nas provas de vinhos.

Um teste não paramétrico (ex: teste de Wilcoxon) teria sido uma escolha melhor pois no teste *t de Student* não há garantia de que os desvios-padrão de um determinado provador de vinho e a média sejam iguais.

5.3. Metodologia usada na avaliação de um painel de provadores de vinhos

5.3.1. Análise com árvore de regressão M5P *Model Trees*

Neste estudo pretende-se verificar se alguns provadores usam uma regra diferente dos outros e, em caso afirmativo, responder às seguintes questões: (1) Em que condições um dado provador X aplica a sua regra diferente do geral e (2) Qual é a forma da regra usada por esse provador e em que aspeto essa difere da regra geral. Para responder a estas questões, usamos os chamados *model trees*, implementados sob o nome M5P, no *software* Weka. Os *model trees* combinam uma árvore de decisão convencional, com a possibilidade de incluir funções de regressão linear nos nós (Frank et al., 2008; Quinlan, 1992).

Neste estudo usamos os dados descritos na secção 5.2.1. com uma única diferença. A variável *Provador* foi representada em forma *numérica*. Em estudos futuros esta variável deverá ser considerada como *nominal*.

5.3.2. Aplicação da árvore de regressão aos dados

Neste estudo utilisámos os dados que foram usados na secção anterior. O algoritmo M5P gerou uma árvore bastante complexa, apresentada no Anexo B4. A seguinte figura mostra um pequeno esboço dessa árvore:

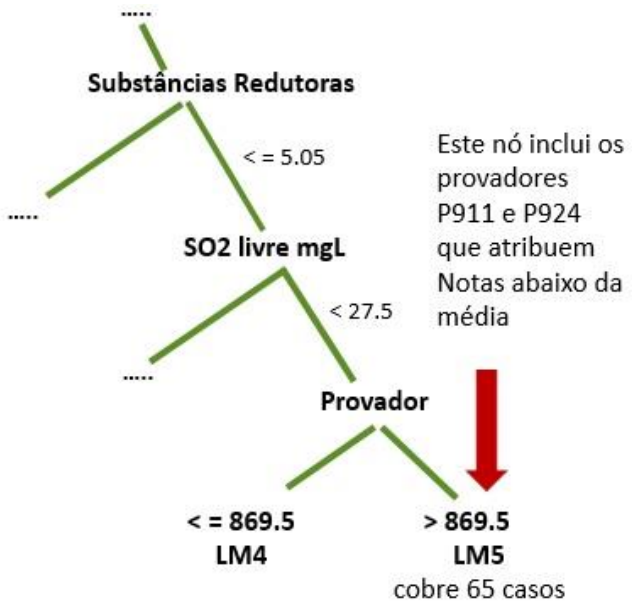

Figura 10 – Árvore de regressão gerada pelo *M5P*

Um dos nós foi atingido com 65 casos que satisfazem as seguintes condições:

- Provador > 869.5;

(inclui provadores que atribuem *Notas* significativamente inferiores à média, i.e. P911 e P924)

- Amostras inferiores a 201.5;
- Ano de Prova superior a 2008.5;
- Substâncias Red menores ou iguais que 5.05
- SO2livremgL inferior ou igual a 27.6.

Este nó cobre 65 provas de vinho de um total de 2268 provas. O comportamento dos provadores é explicado pela regra *LM5*, abaixo apresentado:

Nota =
-0.0058 * Amostra
- 0.0002 * Provador
+ 0.0013 * AnoProva
- 0 * CodVinho
+ 0.0577 * Acidezvolatil-gLacacet
- 0.2686 * pH
+ 0.0811 * AlcoolInf-vv
+ 0.0908 * SO2livre-mgL
- 0.0207 * SO2total-mgL
- 0.0058 * SO2molec-mgL
+ 5.8404 * Densidade-gL
+ 0.0321 * Acideztotal-gLactart
- 4.0417 * ABS420-nm
+ 0.0531 * IPT
+ 0.071 * Extractoseco-gL
+ 0.0001 * Extractonaored-gL
- 0.0015 * Substanciasred
- 0.6547 * Fe-mgL
- 0.2627 * Cu-mgL
+ 3.994

Além disso, um outro nó foi atingido com 48 casos de total de 2268 provas, que satisfazem as seguintes condições:

- Provador <= 768

(este grupo inclui provadores que atribuem Notas significativamente superiores à média, incl. P444 e P467)

- Amostras inferiores a 201.5;
- Ano de Prova superior a 2008.5;
- Substâncias Red maiores que 5.05; CumgL > 0.092
- SO2totalmgL maior do que 117.5.

O comportamento dos provadores é explicado pela regra *LM15*.

Nota =
0.0042 * Amostra
- 0.0021 * Provador
+ 0.0013 * AnoProva
+ 0.0014 * CodVinho
- 1.5436 * Acidezvolatil-gLacacet
- 0.0108 * pH
+ 0.0307 * AlcoolInf-vv
+ 0.0017 * SO2livre-mgL
+ 0.0077 * SO2total-mgL
- 0.019 * SO2molec-mgL
+ 8.7783 * Densidade-gL
- 3.7135 * ABS420-nm
- 0.0018 * IPT
- 0.0006 * Extractoseco-gL
- 0.0699 * Extractonaored-gL
- 0.0015 * Substanciasred
+ 0.0123 * Fe-mgL
- 0.6935 * Cu-mgL
+ 2.8968

O nosso objetivo é comparar as regras acima referidas, LM5 e LM15 com as regras da regressão linear simples, cuja análise será apresentada na próxima secção.

5.4. Análise de regressão linear e comparação com árvores de regressão M5P

Foi efectuada a análise da regressão linear simples com dois objetivos distintos. O primeiro, para obter os coeficientes da regressão linear e, além disso, para obter a informação sobre a importância das variáveis. O segundo, para efectuar um estudo comparativo entre a regressão linear aplicada a todos os dados e as regressões LM5 e LM15 resultantes da aplicação do M5P.

5.4.1. Regressão linear simples

Para gerar a regressão linear a partir de todos os dados usamos o *software* Knime. O resultado obtido apresenta os coeficientes de cada variável e, além disso, os *p-values* que identificam os que mais contribuem para a variável dependente Nota, conforme Tabela 12:

Variáveis	Coef.	Std. Err.	t-value	P>|t|
Cu-mgL	**-3.3383**	**0.5464**	**-6.1099**	**0.0000**
Acideztotal-gLactart	**0.2048**	**0.0359**	**5.7025**	**0.0000**
Acidezvolatil-gLacacet	**3.7267**	**0.6805**	**5.4767**	**0.0000**
ABS420-nm	**-6.6124**	**1.5551**	**-4.2520**	**0.0000**
Alcool-Infvv	0.2407	0.0567	4.2437	0.0000
SO2livre-mgL	**0.0169**	**0.0045**	**3.7508**	**0.0002**
Provador	-0.0004	0.0001	-3.0396	0.0024
Densidade-gL	47.6119	15.9580	2.9836	0.0029
Substanciasred	-0.0599	0.0205	-2.9193	0.0035
SO2molec-mgL	-0.2716	0.1013	-2.6818	0.0074
Fe-mgL	-0.1186	0.0473	-2.5073	0.0122
pH	-0.7736	0.3378	-2.2904	0.0221
Interceção	-35.1541	15.7409	-2.2333	0.0256
Extractoseco-gL	0.0178	0.0111	1.5975	0.1103
IPT	-0.0099	0.0074	-1.3460	0.1784
Extractonaored-gL	-0.0094	0.0169	-0.5529	0.5804
Turbidez-NTU	0.0112	0.0484	0.2306	0.8177
SO2total-mgL	-0.0001	0.0019	-0.0791	0.9369

Tabela 12 - Contribuição para a variável Nota através de regressão linear simples

Conforme verificado no quadro acima, pela análise do *p-value* das variáveis *Cu-mgL*, da Acidez, tanto *Acidezvolatil-gLacacet* como *Acideztotal-gLactart*, entre outras, assinaladas a cor azul, constata-se serem estas que contribuem significativamente para a determinação da Nota. As outras variáveis contribuem pouco.

5.4.2. Comparação entre regressão linear simples e regressões de M5P

Na Tabela 13 foram comparados os coeficientes das três regressões – da regressão geral, da regra LM5 e da regra LM15, com o objetivo de identificar aqueles que apresentam a maior diferença. Nesta tabela foi também introduzida uma coluna rácio que apresenta a diferença entre o valor absoluto maior e o valor absoluto menor. Para isso usamos o seguinte procedimento. Em primeiro lugar convertemos todos os números para valores absolutos. Assim, para o primeiro caso de *Cu-mgL* obtivemos 3.3383, 0.2627 e 0.6935. Depois disso, identificamos o máximo e o mínimo e calculámos o rácio entre os dois. No nosso exemplo obteve 3.3383/0.2627 = 12.71. O rácio foi usado para identificar os coeficientes de variáveis diferentes dos outros.

Variavéis	Coef.RLS	Coef. LM5	Coef.LM15	Rácio
Cu-mgL	-3.3383	-0.2627	-0.6935	**12.71**
Acideztotal-gLactart	0.2048	0.0321	NA	6.38
Acidezvolatil-gLacacet	3.7267	0.0577	-1.5436	**64.59**
ABS420-nm	-6.6124	-4.0417	-3.7135	1.78
AlcoolInf-vv	0.2407	0.0811	0.0307	7.84
SO2livre-mgL	0.0169	0.0908	0.0017	**53.41**
Provador	-0.0004	-0.0002	-0.0021	**10.50**
Densidade-gL	47.6119	5.8404	8.7783	8.15
Substanciasred	-0.0599	-0.0015	-0.0015	**39.93**
SO2molec-mgL	-0.2716	*	*	*
Fe-mgL	-0.1186	*	*	*
pH	-0.7736	*	*	*
Intercept	-35.1541	3.994	2.8968	**12.14**
Extractoseco-gL	0.0178	*	*	*
IPT	-0.0099	*	*	*
Extractonaored-gL	-0.0094	*	*	*
Turbidez-NTU	0.0112	*	*	*
SO2total-mgL	-0.0001	*	*	*

Tabela 13 - Comparação dos coeficientes das três regressões

Os campos da tabela identificados com um asterisco (*) não foram preenchidos por terem sido referidos na tabela 12 como não relevantes para o presente estudo.

Pela análise do quadro acima verifica-se que as variáveis com maior influência na atribuição da Nota (rácio > 10) são *Cu-mgL*, *Acidezvolatil-gLacacet*, *SO2livre-mgL*, *Provador*, *Substanciasred* e *Intercept*.

Como exemplo podemos afirmar que o coeficiente RLS considera que a Nota é 12.71 vezes mais influenciada pela variável *Cu-mgL* do que o coeficiente LM5;

Podemos também dizer que o coeficiente LM5 considera que a Nota é 53.41 vezes mais influenciada pela variável *SO2livre-mgL* do que o coeficiente LM15;

5.4.3. Conclusão do estudo

O principal objetivo deste trabalho consistiu na caraterização dos provadores e sua identificação na contribuição para a atribuição de Notas aos vinhos seleccionados para prova. Para isso, utilizaram-se não só dados estatísticos como também técnicas de *data mining*, tais como, a regressão linear simples e as regressões de M5P. Pela análise estatística identificaram-se 5 provadores que atribuíam Notas diferenciadoras da média, 3 acima da média e 2 abaixo da média. Conhecendo esta informação poderá optar-se por corrigir o viés, positivo ou negativo, das Notas atribuídas pelos provadores às amostras de vinho e/ou eliminar os provadores que não satisfazem os critérios exigidos.

6. Conclusões finais e estudos futuros

O primeiro estudo, efectuado com a CVRVV, tinha por objetivo tentar responder à questão se os vinhos de uma determinada sub-região se assemelhavam a vinhos de outra sub-região. Para isso foram identificados *clusters* com elementos possuindo características semelhantes. O nosso próximo objetivo era de fornecer uma caracterização dos agrupamentos (clusters), pois sem isso, o resultado seria pouco útil ao nosso cliente. Recorremos a várias técnicas para atingir esse objetivo. A primeira consistiu em elaborar o centróide para cada agrupamento (cluster) e compará-los com os centróides dos remanescentes agrupamentos, permitindo identificar, as características físico-químicas e organolépticas que caracterizam cada *cluster*.

Como o número de tais características era bastante grande recorremos a outras técnicas para identificar aquelas que são mais importantes. Em primeiro lugar, usámos a medida de ganho de informação (*information gain*) para isso. Neste estudo utilizamos um agrupamento como o foco, e os elementos de todos os outros agrupamentos foram unidos num único agrupamento de referência. O ganho de informação foi calculado com base nestes dois agrupamentos. Além disso, usamos ainda aprendizagem supervisionada baseada em regras para obter um conjunto de regras que descrevessem as condições diversas para um dado elemento poder ser classificado para o agrupamento de foco. Todas estas técnicas permitiram-nos obter uma boa caracterização dos agrupamentos. Da análise efetuada sobre o conjunto de dados com sub-região identificada formaram-se quatro *clusters* dos quais se destacam dois clusters: um *cluster* composto maioritariamente por vinhos da sub-região de Monção e Melgaço com baixa acidez e teor alcoólico elevado e outro *cluster* composto por vinhos com de 4 sub-regiões (Monção e Melgaço, Amarante, Ave e Cávado) com acidez elevada e de boa qualidade.

Como trabalhos futuros para o estudo da CVRVV, por forma a corroborar os resultados obtidos, pode ser utilizado outro tipo de agrupamento (*clustering*), para analisar as similaridades entre as sub-regiões. Alem disso, podíamos fazer uma análise temporal de um determinado *clusters* ou clusters que permita perceber quais as alterações mais significativas que se verificam, ao longo de um período de tempo, não só nos vinhos de determinada sub-região como também entre sub-regiões.

O segundo estudo, efectuado com a Sogrape Vinhos S.A., consistiu em avaliar um painel de provadores, a partir duma base de dados de provas de vinho, efetuando a análise das notas atribuídas por esses provadores. Pretendeu-se verificar se uns se destacam por um viés

significativo, positivo ou negativo, em relação aos outros provadores. Nos casos em que isso se verificou, pretendeu-se ainda encontrar uma caracterização das condições em que um dado provador sobre- ou sub-avaliou certos vinhos. Além disso, procuramos ainda identificar as características mais importantes do vinho que condicionaram a sua decisão. Essa análise incidiu sobre as regressões lineares geradas por *model trees*, implementados sob o nome M5P em Weka.

O nosso trabalho pode ser usado pela empresa para ou restruturar o painel de avaliadores com a eliminação daqueles provadores que não cumprem com os critérios e cujo viés e variância ultrapasse um dado patamar ou como base para a correção do enviesamento observado.

Como trabalhos futuros nesta área podia-se considerar avaliar a estabilidade do processo adotado. Neste trabalho podíamos recorrer a amostras de dados para verificar se o mesmo resultado pode ser observado em cada amostra. Além disso, podíamos ainda estudar a evolução temporal de cada avaliador. Um outro problema interessante seria verificar se era possível atingir resultados menos enviesados, após os avaliadores serem informados sobre o viés observado depois de cada prova.

7. Referências

Alpuim, J.P. (1997): "Aprendendo a química do vinho". *Química. SPQ* – Soc. Portuguesa Química. Série II. Nº 65.

Braga, R. (2009), *Viticultura de Precisão*. AJAP – Associação dos Jovens Agricultores de Portugal (editores).

Brazdil, P. (2014), "Regression models". Slides. LIAAD-INESC TEC, FEP. Universidade do Porto.

Carvalheira, J. (2011), "A análise sensorial dos vinhos". Laboratório de Química Enológica da DRAPC.http://www.drapc.min-agricultura.pt/base/geral/files/analise_sensorial_vinho_2011.pdf , acedido em 25 Julho 2016.

Cortez P., Teixeira, J., Cerdeira, A., Almeida, F., Matos, T. e Reis, J. (2009a), "Using Data Mining for Wine Quality Assessment". Springer Berlin Heidelberg.

Cortez P., Cerdeira, A., Almeida, F., Matos, T. e Reis, J. (2009b), "Modeling wine preferences by data mining from physicochemical properties". Preprint to Elsevier.

Cortez P., Teixeira, J., Cerdeira, A., Almeida, F., Matos, T. e Reis, J. (2009c), "Wine Quality Datasets". http://www3.dsi.uminho.pt/pcortez/wine/, acedido em 15 Dezembro 2015.

Coutinho, A. (2016), W Aníbal. http://w-anibal.com/glossario, acedido em 28 Julho 2016.

CVRVV (2002), *Catálogo de Marcas da Região dos Vinhos Verdes. A região demarcada dos vinhos verdes – Um século de história*. Comissão de Viticultura da Região dos Vinhos Verdes (CVRVV) (editores).

CVRVV (2005), "ROM - Requisitos organolépticos do Vinho Verde". Comissão de Viticultura da Região dos Vinhos Verdes (CVRVV). http://portal.vinhoverde.pt/pt/documentacao#!, acedido em 2 Junho 2016.

CVRVV (2014), *Um vinho mais fresco – tudo sobre o vinho verde*. Comissão de Viticultura da Região dos Vinhos Verdes (CVRVV) (editores).

CVRVV (2016), "Glossário". Comissão de Viticultura da Região dos Vinhos Verdes (CVRVV). http://portal.vinhoverde.pt/pt/glossario , acedido em 2 Junho 2016.

Daux, V., Cortazar-Atauri, I.G., Yiou, P., Chuine, I.,Garnier, E., Ladurie, E.L.R., Mestre, O. e Tardaguila, J. (2012), "An open-access database of grape harvest dates for climate research- data description and quality assessment". *Climate of the Past*. Vol. 8, pp. 1403-1418.

Dougherty, P.H. (2012), *The Geography of Wine Region*, Springer. Preface. pp. v-vi.

Referências (cont.)

DR (1992), "Estatutos da região demarcada dos vinhos verdes". Decreto-Lei nº10/92 de 3 de Fevereiro. Diário da República I Série-A, Nº28, 3-2-1992. Ministério da Agricultura e do Mar.

DR (2014), Portaria nº216/2014 de 17 de Outubro. Ministério da Agricultura e do Mar.

Frank, E., Wang, Y., Inglis, S., Holmes, G. e Witten, I. (2008), "Using Model Trees for Classification". *Machine Learning*. Vol. 32, pp. 63-76. DOI: 10.1023/A:1007421302149.

Gama, J. (2002), "Cluster Analysis". Slides. LIAAD-INESC TEC, FEP. Universidade do Porto.

Gama, J., Carvalho, A.P.L., Faceli, K., Lorena, A.C. e Oliveira, M. (2012), *Extração de conhecimento de dados – Data Mining*. Edições Sílabo.

Gouveia, C., Liberato, M.L.R. e Dacamara, C.C. (2011), "Modelling past and future wine production in the Portuguese Douro Valley". *Climate Research*. Vol. 48, pp.349-362. Doi: 10.3354/cr01006.

Grainger, K. e Tattersall, H. (2005), *Wine Production: Vine to bottle*. Blackwell Publishing (editors).

Grainger, K. (2009), "Wine quality tasting and selection". John Willey& Sons Publication (editors).

Han, J. e Kamber, M. (2006), *Data Mining. Concepts and Techniques*. Morgan Kaufmann. Elsevier.

Han, J., Kamber, M. e Pei, J. (2012), *Data Mining. Concepts and Techniques*. Morgan Kaufmann. Elsevier.

INE (2015), "Estatísticas Agrícolas 2014". Estatísticas Oficiais. INE.

IVDP (1990), Regulamento (CEE) n.º 2676/90 da Comissão, de 17 de Setembro de 1990. *Instituto dos Vinhos do Douro e do Porto*. http://www.ivdp.pt/pagina.asp?codPag=50&idioma=0&codLei=178, acedido em 11 Abril 2016.

IVDP (2016), Cor. *Instituto dos Vinhos do Douro e do Porto*. https://www.ivdp.pt/pagina.asp?codPag=66, acedido em 28 Julho 2016.

IVV (2009), "Glossário". Instituto da Vinha e do Vinho. http://www.ivv.min-agricultura.pt/np4/155 , acedido em 11 Abril 2016.

Infovini (2016), "Infovini - O portal do vinho português. Glossário". http://www.infovini.com/pagina.php?codNode=18640#tab0 , acedido em 28 Agosto 2016.

Jacobson, J. L. (2006), *Introduction to Wine Laboratory Practices and Procedures*. Springer Science+Business Media, Inc. NY.USA.

James, G., Witten, D. Hastie, T. e Tibshirani, R. (2014), "An Introduction to Statistical Learning with application in R". Springer, New York.

Referências (cont.)

J.O.U.E. (2010), Comunicações oriundas das Instituições, Orgãos e Organismos da União Europeia. Jornal Oficial da União Europeia. Vol. C 43. pp.1-60. http://www.vinhoverde.pt:8081/pt/recursos/documentacao/legislacao/documentos/Comunica%C3%A7%C3%A3o%20%2043%2001%202010%20-%20M%C3%A9todos%20de%20an%C3%A1lise.pdf , acedido em 11 de abril de 2016

Knime software, (2015), Knime Analytics Platform. Knime.com AG. Switzerland.

Marôco, J. (2014). *Análise estatística com o SPSS Statistics*. Editora Report number - análise e gestão de informação, Lda.

Mitchell, T.M. (1997), "Machine Learning". McGraw-Hill Science/Engineering/Math.

Mirkin, B. (2005), *Clustering for Data Mining – A Data Recovering Approach*. Chapman & Hall/CRC.

Parliament of Australia (2001), "Getting a better return – Inquiry into increasing the value added to Australian raw materials". House of Representatives. Standing Committee on Industry, Science and Resources. Commonwealth of Australia. Canberra.

Pestana, M.H. e Gageiro, J.N. (2014), Análise de dados para ciências sociais – A complementaridade do SPSS. Edições Sílabo.

Quinlan, J. (1992). "Learning with continuous class". Procedings AI'92. pp. 343-348. World Scientific. (Adams & Sterling Eds). Singapore.

Ribéreau-Gayon, P.,Glories, Y., Maujean, A. e Dubourdieu, D. (2006). Handbook of Enology – The Chemistry of wine. Stabilization and treatments. Vol 2.

Ribéreau-Gayon, P., Dubourdieu, D., Donèche, B, e Lonvaud, A. (2006a). Handbook of Enology – The Microbiology of Wine and Vinifications. Vol 1. 2nd Edition.

Ribeiro J., Neves, J.M., Machado, J. e Novais, P.J. (2009a), "Wine vinification prediction using data mining tools". ECC'09 Proceedings of the 3rd international conference on European computing conference. *Computing and Computational Intelligence*. WSEAS. pp. 78-85. http://hdl.handle.net/1822/18957 , acedido em 14 de janeiro de 2016.

Ribeiro, J., Neves, J., Sanchez, J., Novais, P. e Machado, J. (2009b), "*Vinification Mining – A Case Study on Wine Production*". Universidade do Minho. http://repositorium.sdum.uminho.pt/handle/1822/18924 , acedido em 14 de janeiro de 2016.

Sallis, P., Shanmuganathan, S., Pavesi, L. e Muñoz, M.C.J. (2008), "Kohonen self-organising maps in the data mining of wine taster comments", *WIT Transactions on Information and Communication Technologies*, Vol. 40, pp.125-139.

Referências (cont.)

Santos, J.A., Grätsch, S.D., Karremann, M.K., Jones, G.V. e Pinto, J.G. (2013), "Ensemble projections for wine production in the Douro Valley of Portugal". *Climatic Change*. Vol. 117, Issue 1, pp. 211–225.

Silva, J.R.M. (2015), "Novas tecnoclogias na gestão da vinha". Diapositivos. *Conferência Internacional da Vinha e do Vinho*. Reguengos de Monsaraz.

Sogrape (2016), Glossário. Sogrape. http://www.sograpevinhos.com/glossario, acedido em 20 de agosto de 2016.

Unwin, T. (2012), "Terroir: At the heart of Geography", in *The Geography of Wine Region*, Dougherty, P.H. (editor), pp. 37-48.

Witten, I. H., Frank, E. e Hall, M. A. (2011), *Data Mining – Practical Machine Learning, Tools and Techniques*. Morgan Kaufmann. Elsevier.

Anexo A1

Características de Identificação (6 atributos)

Os seguintes atributos são identificadores do produto a ser analisado:

Amostra – identifica um determinado vinho, de um ano específico de teste;

Produto – especificação do produto, i.e., vinho verde, aguardente, espumante, vinagre, entre outros;

Tipo – caracteriza o processo de vinificação do vinho, i.e., branco ou tinto;

Casta - Características comuns de um conjunto de videiras, provenientes de uma ou de várias plantas morfologicamente semelhantes, existindo castas autorizadas e recomendadas nas regiões demarcadas. A mesma casta em solos e climas diferentes origina vinhos diferenciados, embora algumas componentes aromáticas próprias da casta se mantenham (Sogrape, 2016).

Região – identifica a sub-região da Região Demarcada de Vinho Verde (RDVV). Segundo a CVRVV, é constituída por uma área geográfica bem definida, com vinhas inscritas para a produção de vinho de qualidade (DR, 1992).

Tipo I – engloba todas as denominações especiais que uma colheita pode ter, i.e, reserva especial, colheita seleccionada, garrafeira, entre outras.

Características físico-químicas (25 atributos)

Os valores destes atributos são determinados através de ensaios laboratoriais e servem de controlo e certificação dos vinhos pelas entidades competentes, neste caso, a CVRVV.

Acidez Fixa (AcidFix) – resultado da diferença entre a *acidez total* e a *acidez volátil*, tendo como acidez fixa mínima exigida 4,5 g/L ácido tartárico (CVRVV, 2016).

Acidez Total (AcidTot) – Soma dos ácidos tituláveis, quando se eleva o vinho a pH7, pela adição de uma solução alcalina titulável (o CO2 e o SO2 total não estão compreendidos neste valor) (CVRVV, 2016). Segundo os autores (Ribéreau-Gayon et al., 2006) fazem parte dos ácidos tituláveis o ácido tartárico, málico, cítrico, láctico, succínico e acético, excluindo o ácido carbónico e o dióxido de enxofre.

Acidez Volátil* (*AcidVolat) - Componente presente no vinho que, em dose elevada, origina o aroma a vinagre. Em excesso é o resultado da falta de cuidados durante a vinificação. Nos vinhos velhos é habitual um toque mais acentuado de acidez volátil, aos quais se dá a designação de "vinagrinho" (CVRVV, 2016; Sogrape, 2016). Os autores (Ribéreau-Gayon et al., 2006) dizem ser esta a variável fisico-química mais importante a ser monitorada no processo de produção do vinho. Faz parte integral da Acidez Total mas é considerada à parte, estando fortemente relacionada com a qualidade do vinho. A sua presença em excesso num vinho é considerado, por um enólogo, como um factor negativo.

Acido Ascórbico (AcidAsc) - Ou vitamina C, quando adicionado ao mosto durante a vinificação, juntamente com o dióxido de enxofre, impede a oxidação e ajuda a manter frescos os vinhos brancos (Sogrape, 2016).

Anexo A1(cont.)

Acido Cítrico (AcidCitric) - Ácido constitutivo dos vinhos que proporciona acidez fresca. Por vezes, pode ser atacado pelas bactérias da fermentação maloláctica (Sogrape, 2016).

Acido Sórbico (AcidSorb) - Aditivo muito utilizado nas indústrias alimentar e de bebidas para neutralizar leveduras e bolores. Cheira excessivamente a folhas de gerânio pisadas para quem é muito sensível (Sogrape, 2016).

Açucares Redutores (AcucarRed) - Componente presente no vinho que, em dose elevada, origina o aroma a vinagre. Em excesso é o resultado da falta de cuidados durante a vinificação. Nos vinhos velhos é habitual um toque mais acentuado de acidez volátil, aos quais se dá a designação de "vinagrinho"(CVRVV, 2016). Segundo Ribéreau-Gayon Et al. (2006), estes açúcares são compostos por pentoses e hexoses. As hexoses (glicose e frutose), são açúcares fermentescíveis, utilizados como alimento pelas leveduras, precursores diretos do etanol, mas também podem ser consumidos por bactérias, e as pentoses (arabinose e xilose), não são fermentáveis.

Açucares Totais (AcucarTot) – componentes essenciais do mosto que se transformam em álcool e em outras substâncias por acção das leveduras (Sogrape, 2016).

Cloretos (Cloret) – ensaio laboratorial para detectar a existência de vários tipos de cloretos. A sua concentração é inferior a 50 mg/l, e exprime-se em cloreto de sódio (Ribéreau-Gayon Et al.; 2006). Segundo Alpuim (1997) vinhos com elevada presença deste componente (NaCl) pode significar como sendo originários de solos salgados.

Cobre - Odor desagradável de um vinho alterado e estragado pela presença de cobre. O excesso de cobre (mais de 1 mg/l) detecta-se imediatamente nos vinhos brancos devido à sua cor parda (CVRVV, 2016).

Dióxido Enxofre Livre (DioxEnxLiv) – análise para identificação desta substância química. A sua presença "livre" é justificada por o dióxido de enxofre ainda não ter sido combinado, quando foi adicionado ao mosto na pré-fermentação, mas que poderá acontecer durante o período da fermentação (Ribéreau-Gayon Et al.; 2006a).

Dioxido Enxofre Total (DioxEnxTot) – composto por dióxido de enxofre livre e por dióxido de enxofre combinado. O dióxido de enxofre ou anidrido sulfuroso - SO_2 – é uma substância desinfectante que se emprega para garantir o controlo e a limpeza na elaboração de vinhos. Tem propriedades clarificantes, antioxidantes e anti-sépticas (Sogrape, 2016). O seu emprego está estritamente regulamentado (SO2 total max:210 ou max: 260 se açúcares redutores >=5) pelo Regulamento CE 1493/1999 de 17 de Maio, conforme informação referida no ficheiro fornecido pela CVRVV.

Extrato Não Redutor (ExtrNRed) – é o extrato seco total diminuído dos açúcares totais (IVDP, 1990).

Extrato Seco Total (ExtrSecTot) – respeitante às matérias secas totais; conjunto de todas as substâncias que, em determinadas condições de temperatura e pressão, não se volatilizam. Exprime-se em gramas por litro. Um vinho com pouco extrato é leve; com muito é espesso na boca (IVDP, 1990).

Massa Volúmica (MassVol), ou Densidade - é o quociente entre a massa de um determinado *volume de vinho* ou de mosto a 20 °C e esse volume. Exprime-se em gramas por mililitro e o seu símbolo é 20 °C. A sua determinação apresenta algumas vantagens quanto à concentração de açúcares existentes nos mostos (IVDP, 1990). Jacobson (2006) refere que Hagglung define a densidade e a gravidade específica como propriedades físicas de uma substância. A densidade é definida como a massa por volume de uma substância e a gravidade específica é o ratio da densidade de uma substância medida a uma determinada temperatura face à densidade de um material de referência, normalmente a água, também medida a uma determinada temperatura. A densidade e gravidade específica do vinho é menor do que a da água. Quanto mais álcool estiver presente no vinho menor a sua densidade pois o álcool tem uma densidade menor do que a água. Quantos mais sólidos ou materiais insolúveis estiverem presentes no vinho maior a densidade pois tornam o vinho mais denso.

pH - Potencial de hidrogénio. Índice de acidez ou de alcalinidade de um vinho. Calcula-se pelo co-logaritmo de concentração em iões hidrogénio. Quando o valor em pH é inferior a 7, o líquido é ácido. Os solos da vinha distinguem-se também pelo seu pH maior ou menor. O pH de um mosto varia entre 2,8 a 3,8; o pH de um vinho, varia de 3 a 4 (CVRVV, 2016).

Relação Alcool/Peso e Extrato Redutor (RelAlcoolPeso_ExtrNRedutor) - ensaio laboratorial para verificar a relação entre álcool / peso e extrato redutor.

Sobrepressão (Sobrpress) – Segundo o IVV (2009), existem muitos tipos de prensas (verticais, horizontais, pneumáticas, hidropneumáticas) que se utilizam para extrair o mosto na vinificação em brancos ou para a obtenção de vinho de prensa na vinificação em tintos. Ao prensar o vinho é importante medir bem a pressão aplicada, para não martirizar a vindima nem extrair sabores herbáceos e óleos essenciais das peles e das grainhas das uvas. Este termo também pode designar o vinho elaborado com o produto da prensagem dos engaços. Uma parte do vinho de prensa é acrescentada ao vinho de gota, segundo o critério do produtor. Designa-se prensagem à operação que consiste na separação das matérias sólidas de uma vindima antes ou após a fermentação.

Sulfatos (Sulfat) – ensaio laboratorial para verificar a presença de um sal cuproso, denominado sulfato de cobre, que se utiliza como fungicida na vinha (Sograpе, 2016).

Título Alcoométrico Volúmico Adquirido (TitAlcVolAdq) – ensaio laboratorial para determinar a graduação alcoólica do vinho. O seu intervalo de variação é entre os 8° e os 11,5°, mas no caso particular do vinho verde casta Alvarinho o mínimo é 11,5, conforme informação referida no ficheiro fornecido pela CVRVV. Devido à baixa densidade do etanol, os vinhos secos, que contêm valores negligenciáveis de açúcar, possuem uma densidade inferior à da água (1.00), variando entre 0.91 e 0.94. Este valor diminui à medida que a percentagem de álcool aumenta (Ribéreau-Gayon Et al.; 2006).

Título Alcoométrico Volúmico Total (TitAlcVolTot) – soma do título alcoométrico volúmico adquirido e do álcool potencial (açucares residuais). O seu intervalo de variação é entre os 8,5° e os 14°, conforme informação referida no ficheiro fornecido pela CVRVV.

Título Alcoométrico Volúmico Total Registado (TitAlcVolTotReg) – soma do título alcoométrico volúmico adquirido e do álcool potencial (açucares residuais) para verificação da conformidade da amostra com o padrão registado no CVRVV. O seu intervalo de variação é entre os 8,5° e os 14° (CVRVV).

Metanol - A presença de metanol é traduzida por uma diminuição do índice de refracção e, por conseguinte, do título alcoométrico, conforme refere o Regulamento (CEE) n.º 2676/90 da Comissão, de 17 de Setembro de 1990 (IVDP, 1990).

Características Organolépticas (8 atributos)

Os seguintes atributos são determinados pelo enólogo:

Aspeto Limpidez (AspLimpid) - Diz-se da cor, aroma ou sabor de um vinho bem elaborado, sem sedimentos nem alterações. Os vinhos sadios são necessariamente límpidos, mas certos vinhos, em especial os que envelhecem muitos anos, podem apresentar sedimento que fica no fundo da garrafa em repouso e não altera a sua limpidez (CVRVV, 2016). No presente estudo o valor mínimo registado é 1 e o valor máximo é 5 em que, segundo o art. 11° dos Estatutos da Região Demarcada dos Vinhos Verdes, o valor mínimo é límpido e o valor máximo é ligeiramente opalino.

Aspecto Cor (AspCor) - Aspecto cromático do vinho. Há vinhos brancos, rosés e tintos, consoante a sua cor. A intensidade da cor deriva das castas, do clima e do método de vinificação (CVRVV, 2016). Segundo os autores Ribéreau-Gayon et al. (2006) a atribuição da cor aos vinhos brancos é muito mais complexa do que aos outros vinhos pois o espetro não tem definido visivelmente o seu valor máximo. A absorção é contínua entre 500 e 280 nm, com um máximo de abertura para os raios ultra violetas. No presente estudo o valor mínimo registado é 1 e o valor máximo é 6 em que, segundo o art. 11° dos Estatutos da Região Demarcada dos Vinhos Verdes, o valor mínimo é citrino descorado e o valor máximo é ligeiramente dourado.

Aroma Defeito Marcado (AromaDefMarc) – verificação de ausência de defeito marcado no aroma. No presente estudo o valor mínimo registado é 1 e o valor máximo é de 2. Conforme art. 11° dos Estatutos da Região Demarcada dos Vinhos Verdes, o valor mínimo é 2, referente a "Não" (tem defeito marcado no aroma).

Aroma Qualidade (AromaQualid) - verificação da qualidade do aroma (notação igual ou superior a 5) para ser classificado como vinho verde, segundo escala abaixo representada na Figura 11. No caso do vinho verde de casta deve cumprir os requisitos de Vinho Verde, evidenciar a casta e ter uma notação igual ou superior a 6 (CVRVV, 2005). O Vinho Verde com indicação de sub-região ou com designativo de qualidade deve cumprir os requisitos de Vinho Verde e apresentar características organolépticas destacadas, com notação superior ou

igual a sete para os designativos Superior e Colheita Seleccionada, de acordo com a escala de qualidade acima referida e abaixo representada na Figura 11.

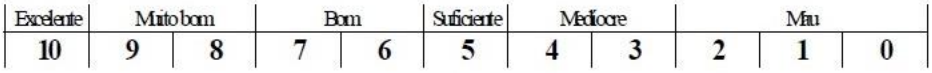

Figura 111 - Escala de Qualidade

Sabor Qualidade (SaborQualid) **-** verificação de qualidade suficiente (notação igual ou superior a 5) para ser classificado como vinho verde. No caso do vinho verde de casta deve cumprir os requisitos de vinho verde, evidenciar a casta e ter uma notação igual ou superior a 6 (CVRVV, 2005). Esta variável é medida segundo a escala da Figura 11 acima representada (CVRVV, 2005).

Aroma Tipicidade (AromaTipic) – verificação da tipicidade (notação igual ou superior a 5) para ser classificado como vinho verde, segundo escala abaixo representada na Figura 12 (CVRVV, 2005).

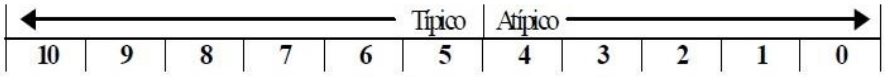

Figura 12 - Escala de Tipicidade

Sabor Tipicidade (SaborTipic) **-** verificação da tipicidade (notação igual ou superior a 5) para ser classificado como vinho verde, segundo escala da Figura 12 acima representada (CVRVV, 2005).

Sabor Defeito Marcado (SaborDefMarc) - verificação de ausência de defeito marcado no sabor. No presente estudo o valor mínimo registado é 1 e o valor máximo é de 2. Conforme art. 11° dos Estatutos da Região Demarcada dos Vinhos Verdes, o valor mínimo é 2, referente a Não (tem defeito marcado no sabor).

Anexo A2

	Variável	Descrição	Codificação	Tipo de
Dados identificadores do produto	Amostra	Identifica a prova		Discreta
	CASTA	Característica comum a conjunto de videiras	Alvarinho	Categórica
			Arinto	
			Arinto / Pedernã	
			Avesso	
			Azal	
			Batoca	
			Fernão Pires	
			Fernão Pires / Maria Gomes	
			Loureiro	
			Trajadura	
			Sem casta	
	REGIAO	Sub-região da RDVV	Amarante	Categórica
			Ave	
			Baião	
			Basto	
			Cávado	
			Lima	
			Monção e Melgaço	
			Paiva	
			Sousa	
			Sem Região	
	TIPOI	Denominações especiais	Colheira seleccionada	Categórica
			Escolha	
			Grande escolha	
			Reserva	
			Superior	
			Vindima tardia	
			Sem tipo	
Características Físico químicas	AcidFix	Acidez Fixa	3 a 13.9	Contínua
	AcidTot	Acidez Total	3.2 a 14.2	Contínua
	AcidVolat	Acidez Volátil	0.22 a 1.33	Contínua
	AcidCitric	Ácido Cítrico	-0.04 a 2.2	Contínua
	Cloret	Cloretos	0.006 a 0.65	Contínua
	DioxEnxLiv	Dióxido de Enxofre Livre	0 a 367	Contínua
	DioxEnxTot	Dióxido de Enxofre Total	8 a 507	Contínua
	ExtrNRed	Extrato Não Redutor	11.9 a 34.7	Contínua
	ExtrSecTot	Extrato Seco Total	14.6 a 162	Contínua
	MassVol	Massa Volúmica	0.98644 a 1.04494	Contínua
	pH	pH	2.53 a 3.87	Contínua
	Sulfat	Sulfatos	0.16 a 1.97	Contínua
	TitAlcVolAdq	Titulo Alcoométrico Volúmico Adquirido	6.5 a 14.9	Contínua
Características Organolépticas	AspLimpid	Aspeto Limpidez	1 a 5	Discreta
	AspCor	Aspeto Cor	1 a 6	Discreta
	AromaDefMarc	Aroma Defeito Marcado	1 a 2	Discreta
	AromaQualid	Aroma Qualidade	2 a 9	Discreta
	AromaTipic	Aroma Tipicidade	4 a 9	Discreta
	SaborDefMarc	Sabor Defeito Marcado	1 a 2	Discreta
	SaborQualid	Sabor Qualidade	2 a 9	Discreta
	SaborTipic	Sabor Tipicidade	4 a 9	Discreta

Anexo A3

Conforme referido na secção 4.2.7. vão apresentar-se aqui os remanescentes *clusters* do DadosII: 1, 2, 3, 4, 5, 6, 7, 8 e 11.

Cluster 1 – Vinhos com Densidade e Dióxido de Enxofre Total elevados

Características: 1208 elementos, que representam 8.4 % de dados.
A regra cobre 1141 amostras e estão correctamente cobertos 1045 amostras.

	Cluster 1	Características	InfoGain	Centróide	Centróide de Outros Clusters	Regra
1	MassVol	Fis.-Quím.	0.454	1.263	-0.116	MassVol > 0.43
2	DioxEnxTot	Fis.-Quím.	0.449	1.526	-0.140	DioxEnxTot > 0.32
3	TitAlcVolAdq	Fis.-Quím.	0.346	-1.106	0.102	TitAlcVolAdq <= 0.35
4	ExtrSecTot	Fis.-Quím.	0.311	1.001	-0.092	ExtrSecTot > -0.39
5	SaborQualid	Sens.	0.284	5.325	6.156	SaborQualid <= 6.5
6	AromaQualid	Sens.	0.283	5.327	6.158	
7	TitAlcVolTot	Fis.-Quím.	0.257	-0.907	0.083	
8	AromaTipic	Sens.	0.235	5.448	6.209	
9	SaborTipic	Sens.	0.234	5.447	6.209	
10	AcucarRed	Fis.-Quím.	0.201	0.706	-0.065	AcucarRed <= 4.44
11	RelAlcPeso_ExtNRed	Fis.-Quím.	0.194	-0.635	0.058	RelAlcPeso_ExtNRed <= 0.98
12	DioxEnxLiv	Fis.-Quím.	0.189	0.962	-0.088	
13	Cloret	Fis.-Quím.	0.183	0.095	-0.009	Cloret <= 3.71
14	REGIAO	Identificação	0.139			
15	CASTA	Identificação	0.123			
16	AcidTot	Fis.-Quím.	0.072	0.338	-0.031	AcidTot > -1.56
17	AcidFix	Fis.-Quím.	0.070	0.315	-0.029	
18	AcidCitric	Fis.-Quím.	0.070	0.621	-0.057	AcidCitric > -1.46
19	TIPOI	Identificação	0.068			
20	Sulfat	Fis.-Quím.	0.068	0.467	-0.043	
21	AcucarTot	Fis.-Quím.	0.063	0.186	-0.017	

Este agrupamento parece estar caracterizado por:

- Valor alto de ***Massa Volúmica***, *MassVol* (InfoGain=0.454). A regra estipula que *MassVol* > *0.43* e o valor deste atributo no centróide *MassVol=1.263* ultrapassa largamente o valor exigido. O valor dos outros *clusters* é bem mais baixo (-0.116).

- Valor alto de ***Dióxido de Enxofre Total***, *DioxEnxTot* (InfoGain=0.449). A regra estipula que *DioxEnxTot* > -0.92 e *DioxEnxTot* < 0.32 e o valor deste atributo no centróide é *DioxEnxTot = 1.526*. O valor doutros clusters é bastante inferior (-1.140).

- Valor baixo de ***Título Alcoométrico Volúmico Adquirido***, *TitAlcVolAdq* (InfoGain=0.346). A regra estipula que *TitAlcVolAdq <= 0.35* e o valor deste atributo no centróide é *TitAlcVolAdq=-1.106*. O valor doutros *clusters* é bastante mais elevado (0.102).

- Valor baixo de ***Extrato Seco Total***, ExtrSecTot (InfoGain=0.311). A regra estipula que *ExtrSecTot > -0.39* e o valor deste atributo no centróide é ExtrSecTot=*1.001*. O valor dos centróides de outros clusters é mais reduzido (-0.092).

- Valor baixo de ***Sabor Qualidade***, SaborQualid (InfoGain=0.284). A regra estipula que *SaborQualid <= 6.5* e o valor deste atributo no centróide é *SaborQualid*=5.325 e no centróide dos outros *clusters* regista valor superior, 6.156.

- Valor baixo de ***Açucares Redutores***, *AcucarRed* (InfoGain=0.201). A regra estipula que *AcucarRed <= 4.44* e o valor deste atributo tanto neste centróide como nos outros é significativamente baixo, *AcucarRed =0.706* e *AcucarRed =-0.065*, respectivamente.

- Valor baixo de ***Relação AlcoolPeso_Extrato Não Redutor***, *RelAlcPeso_ExtNRed* (InfoGain=0.194). A regra estipula que *RelAlcPeso_ExtNRed <= 0.98* e o valor deste atributo no centróide é significativamente baixo, *RelAlcPeso_ExtNRed =-0.635*. O valor dos centróides de outros *clusters* é bastante superior (0.058).

- Valor alto de **Cloretos**, *Cloret* (InfoGain=0.189). A regra indica *Cloret* <= 3.71 e o seu valor no centróide é *Cloret =0.095*. O valor dos centróides de outros *clusters* é inferior (-0.009).

Observação (hipótese):

Possui ***Massa Volúmica*** elevada, sinónimo de maior presença ou de açucares ou de ácidos, aliado ao excesso de ***Dióxido de Enxofre Total***, com média de 184.05, quando a legislação estabelece um valor máximo de 210, adquirindo o vinho um aroma picante e um gosto final desagradável, e a um baixo ***Título Alcoométrico Volúmico Adquirido***, aparado a 5% ou não aparado, de 9.59°/9.57°, respectivamente, (o valor mínimo para a RDVV é 8°) comprovando não registar outliers significativos podendo servir como indicadores do tipo de vinhos deste cluster. A maior saliência destas características pode indicar o baixo valor médio de 5.32, quando o valor mínimo aceitável é de 5 pontos, atribuído *ao* ***Sabor Qualidade***.

Este *cluster* é constituído maioritariamente por vinhos sem sub-região identificada (95%), sem presença significativa de nenhuma sub-região, como pode ser constatado pela distribuição das amostras por regiões na seguinte tabela:

Sub-regiões	Nr amostras	Amostras %	c1	c1 %
Amarante	501	3.49	22	1.82
Ave	471	3.28	9	0.75
Baiao	320	2.23	5	0.41
Basto	391	2.72	7	0.58
Cavado	431	3.00	7	0.58
Lima	465	3.24	4	0.33
Monção e Melgaço	1808	12.59	2	0.17
Paiva	155	1.08	0	0.00
Sousa	399	2.78	4	0.33
S/região	9414	65.58	1148	**95.03**
Totais	14355	100	1208	100

Cluster 2 – Vinhos com acidez elevada

Características: 1876 elementos, que representam 13.1 % de dados.
A regra cobre 1909 amostras e estão correctamente cobertos 1779 amostras.

	Cluster 2	Características	InfoGain	Centróide	Centróide de Outros Clusters	Regra
1	AcidTot	Fis.-Quím.	0.631	1.635	0.043	
2	AcidFix	Fis.-Quím.	0.628	1.648	0.031	AcidFix > -0.56
3	ExtrNRed	Fis.-Quím.	0.177	0.794	0.039	
4	pH	Fis.-Quím.	0.119	-0.710	-0.079	
5	TitAlcVolAdq	Fis.-Quím.	0.086	0.118	-0.252	
6	AcidCitric	Fis.-Quím.	0.066	0.228	0.172	
7	MassVol	Fis.-Quím.	0.060	0.045	0.121	MassVol > -1.51
8	Cloret	Fis.-Quím.	0.058	-0.210	0.486	
9	SaborTipic	Sens.	0.053	7.094	6.209	SaborTipic > 6.5
10	AromaTipic	Sens.	0.052	7.095	6.210	
11	SaborQualid	Sens.	0.050	7.082	6.156	
12	AromaQualid	Sens.	0.050	7.085	6.158	
15	REGIAO	Identificação	0.037			
17	CASTA	Identificação	0.029			
33	TIPOI	Identificação	0.005			

Este agrupamento parece estar caracterizado por:

- Valor alto de ***Acidez Fixa***, *AcidFix* (InfoGain=0.628). A regra estipula que *Acidez Fixa* > *-0.56* e o valor deste atributo no centróide é *Acidez Fixa =1.648*.

Observação (hipótese):
O item *valor de* ***Acidez Fixa*** sugere que este grupo inclui vinhos bastante ácidos. A sua média, não aparada ou aparada a 5% tem o mesmo valor de 6.64, denotando não haveroutliers extremos. Dado que o mínimo estabelecido pelos Estatutos da RDVV é de 5.4, corrobora assim a hipótese acima avançada.
Este cluster é constituído maioritariamente por vinhos sem sub-região identificada (51%), e temos também vinhos de Monção e Melgaço (13%), vinhos do Cávado (7%), vinhos do Lima (7%) e do Sousa (5%), como pode ser constatado pela distribuição das amostras por regiões na seguinte tabela:

Sub-regiões	Nr amostras	Amostras %	c2	c2%
Amarante	501	3.49	64	3.41
Ave	471	3.28	92	4.90
Baiao	320	2.23	61	3.25
Basto	391	2.72	53	2.83
Cavado	431	3.00	137	7.30
Lima	465	3.24	133	7.09
Monção e Melgaço	1808	12.59	249	13.27
Paiva	155	1.08	23	1.23
Sousa	399	2.78	103	5.49
S/região	9414	65.58	961	51.23
Totais	14355	100	1876	100

Cluster 3 – Vinhos com baixo teor alcoólico

Características: 2230 elementos, que representam 15.5 % de dados.
A regra cobre 2176 amostras e estão correctamente cobertos 2008 amostras.

	Cluster 3	Características	InfoGain	Centróide	Centróide de Outros Clusters	Regra
1	SaborTipic	Sens.	0.325	6.018	6.168	
2	AromaTipic	Sens.	0.323	6.02	6.169	AromaTipic > 5.5
3	SaborQualid	Sens.	0.316	5.988	6.104	SaborQualid <= 6.5
4	AromaQualid	Sens.	0.316	5.99	6.106	
5	MassVol	Fis.-Quím.	0.201	0.067	0.095	-1.11 < MassVol <= 1.68
6	TitAlcVolAdq	Fis.-Quím.	0.174	-0.304	-0.186	-2.03 < TitAlcVolAdq <= 1.75
7	AcidTot	Fis.-Quím.	0.148	-0.556	0.147	AcidTot <= 0.82
8	AcidFix	Fis.-Quím.	0.143	-0.500	0.131	
9	TitAlcVolTot	Fis.-Quím.	0.115	-0.342	-0.183	TitAlcVolTot > -2.10
10	pH	Fis.-Quím.	0.090	0.512	-0.138	pH > -2.32
11	CASTA	Identificação	0.052			
13	REGIAO	Identificação	0.044			
19	TIPOI	Identificação	0.024			

Este agrupamento parece estar caracterizado por:

- Valor baixo de ***Aroma Tipicidade***, *AromaTipic* (*InfoGain*=0.323). A regra estipula que *AromaTipic* > 5.5 e o valor deste atributo no centróide é significativamente baixo, *AromaTipic* =6.02, enquanto que nos outros centróides é de 6.169.

- Valor baixo de ***Sabor Qualidade***, *SaborQualid* (*InfoGain*=0.316). A regra estipula que *SaborQualid* <= 6.5 e o valor deste atributo no centróide é *SaborQualid* =5.988 enquanto nos centródes dos outros *clusters* é de 6.104.

- Valor de ***Massa Volúmica***, *MassVol* (*InfoGain*=0.201). A regra estipula que -1.11 < *MassVol* <= 1.68 e o valor deste atributo no centróide é um pouco acima desta média, *MassVol* =0.067.

- Valor baixo de ***Título Alcoométrico Volúmico Adquirido***, *TitAlcVolAdq* (*InfoGain*=0.174). A regra estipula que -2.03 < *TitAlcVolAdq* <= 1.75 e o valor deste atributo no centróide é *TitAlcVolAdq*=-0.304.

- Valor baixo de ***Título Alcoométrico Volúmico Total***, *TitAlcVolTot* (*InfoGain*=0.116). A regra estipula que *TitAlcVolTot* <= 1.11 e o valor deste atributo no centróide é *TitAlcVolTot*=-0.366.

- Valor de ***Acidez Total*** baixo, *AcidTot* (*InfoGain*=0.148). A regra estipula que *AcidTot* <= 0.82 e o valor deste atributo no centróide é *AcidTot* =-0.556.

Observação (hipótese):

Estes vinhos apresentam um baixo valor no item ***Aroma Tipicidade***, denotando pouca relação com os vinhos verdes. No entanto, dado que este valor é organoléptico, a sua média, não aparada ou aparada a 5%, é de 6.02 pontos (Escala de 0 a 10, sendo 5 o valor mais baixo a identificar tipicidade). A corroborar temos os valores dos ***Título Alcoométrico Volúmico Adquirido*** e ***Título Alcoométrico Volúmico Total*** baixos apontam que este grupo é constituído por vinhos

com baixo teor alcoólico. Acresce o baixo valor de ***pH***, com média, não aparada ou aparada a 5%, de 3.26, quando o *pH* de um vinho varia entre 3 e 4, demonstrando estarmos em presença de vinhos bastante ácidos.

Este *cluster* é constituído maioritariamente por vinhos sem sub-região identificada (81%), sem presença significativa de nenhuma sub-região, como pode ser constatado pela distribuição das amostras por regiões na seguinte tabela:

Sub-regiões	Nr amostras	Amostras %	c3	c3%
Amarante	501	3.49	35	1.57
Ave	471	3.28	74	3.32
Baiao	320	2.23	10	0.45
Basto	391	2.72	34	1.52
Cavado	431	3.00	64	2.87
Lima	465	3.24	57	2.56
Monção e Melgaco	1808	12.59	78	**3.50**
Paiva	155	1.08	20	0.90
Sousa	399	2.78	54	2.42
S/região	9414	65.58	1804	**80.90**
Totais	14355	100	2230	100

Cluster 4 – Vinhos com Densidade e Extrato Seco Total elevados

Características: 1574 elementos, que representa 11% de dados.
A regra cobre 1243 amostras e estão correctamente cobertos 1202 amostras.

	Cluster 4	Características	InfoGain	Centróide	Centróide de Outros Clusters	Regra
1	MassVol	Fis.-Quím.	0.504	1.308	-0.029	MassVol > 0.57
2	ExtrSecTot	Fis.-Quím.	0.403	1.135	-0.121	ExtrSecTot > -0.38
3	AcucarRed	Fis.-Quím.	0.366	0.885	-0.131	AcucarRed > -0.26
4	TitAlcVolAdq	Fis.-Quím.	0.290	-1.004	-0.116	TitAlcVolAdq <= 1.05
5	AcucarTot	Fis.-Quím.	0.249	0.603	-0.051	AcucarTot > -0.01
6	TitAlcVolTot	Fis.-Quím.	0.160	-0.369	-0.180	
7	pH	Fis.-Quím.	0.147	-0.751	-0.011	pH <= 1.02
8	ExtrNRed	Fis.-Quím.	0.145	-0.703	0.137	ExtrNRed <= 1.21
9	TitAlcVolTotReg	Fis.-Quím.	0.143	-0.289	0.005	
10	AromaQualid	Sens.	0.133	5.972	6.088	AromaQualid > 4.5
11	SaborQualid	Sens.	0.131	5.97	6.086	SaborQualid <= 7.5
12	Cloret	Fis.-Quím.	0.129	0.104	0.497	Cloret <= 3.49
13	REGIAO	Identificação	0.128			
14	SaborTipic	Sens.	0.127	6.008	6.145	SaborTipic > 5.5
15	AromaTipic	Sens.	0.126	6.009	6.146	
16	CASTA	Identificação	0.114			
17	DioxEnxTot	Fis.-Quím.	0.111	0.487	-0.028	-2.01 < DioxEnxTot <= 2.48
18	AspCor	Sens.	0.097	1.945	2.201	AspCor <= 0.59
19	Metanol	Fis.-Quím.	0.096	-0.207	0.012	Metanol <= 1.91
20	AcidSorb	Fis.-Quím.	0.092	0.185	-0.032	
21	Cobre	Fis.-Quím.	0.073	0.053	0.006	
22	DioxEnxLiv	Fis.-Quím.	0.072	0.429	-0.066	DioxEnxLiv <= 4.43
23	Sobrepress	Fis.-Quím.	0.066	0.049	-0.011	
24	RelAlcPeso_ExtNRed	Fis.-Quím.	0.055	0.123	-0.144	RelAlcPeso_ExtNRed > -0.92
25	AcidTot	Fis.-Quím.	0.055	-0.067	0.098	AcidTot <= 2.13
26	AcidFix	Fis.-Quím.	0.053	-0.012	0.082	
27	TIPOI	Identificação	0.046			

Este agrupamento parece estar caracterizado por:

- Valor de ***Massa Volúmica*** positivo, *MassVol* (*InfoGain*=0.503). A regra estipula que *MassVol* > 0.57 e o valor deste atributo no centróide é *MassVol* =1.308.

- Valor alto de ***Extrato Seco Total***, *ExtrSecTot* (InfoGain=0.403). A regra estipula que *ExtrSecTot* > -0.38 e o valor deste atributo no centróide é *ExtrSecTot* = 1.135.

- Valor positivo mas baixo de ***Açucares Redutores***, *AcucarRed* (*InfoGain*=0.366). A regra estipula que *AcucarRed* > -0.26 e o valor deste atributo no centróide é *AcucarRed* =0.885.

- Valor baixo de ***Título Alcoométrico Volúmico Adquirido***, *TitAlcVolAdq* (*InfoGain*=0.290). A regra estipula que *TitAlcVolAdq* <= 1.05 e o valor deste atributo no centróide é *TitAlcVolAdq*=-1.004.

- Valor alto de ***Açucares Totais***, *AcucarTot* (*InfoGain*=0.249). A regra estipula que *AcucarTot* > -0.01 e o valor deste atributo no centróide é *AcucarTot* =0.603.

- Valor baixo de ***pH***, *pH* (*InfoGain*=0.147). A regra estipula que *pH* <= 1.02 e o valor deste atributo no centróide é *pH* =-0.751.

- Valor baixo de ***Extrato Não Redutor***, *ExtrNRed* (*InfoGain*=0.145). A regra estipula que *ExtrNRed* <= 1. 21 e o valor deste atributo no centróide é *ExtrNRed* =-0.703.

- Valor baixo de ***Aroma Qualidade***, *AromaQualid* (*InfoGain*=0.133). A regra estipula que *AromaQualid* > 4.5 e o valor deste atributo no centróide é significativamente baixo, *AromaQualid* =5.972, mesmo nos centróides dos outros clusters, 6.088.

- Valor baixo de ***Sabor Qualidade***, *SaborQualid* (*InfoGain*=0.131). A regra estipula que *SaborQualid* <= 7.5 e o valor deste atributo no centróide é *SaborQualid* =5.97.

- Valor baixo de ***Cloretos***, *Cloret* (*InfoGain*=0.129). A regra estipula que *Cloret* <= 3.49 e o valor deste atributo no centróide é relativamente baixo, *Cloret*=0.104.

- Valor baixo de ***Sabor Tipicidade***, *SaborTipic* (*InfoGain*=0.127). A regra estipula que *SaborTipic* > 5.5 e o valor deste atributo no centróide é significativamente baixo, *SaborTipic* =6.008.

- Valor positivo de ***Dióxido de Enxofre Total***, *DioxEnxTot* (InfoGain=0.111). A regra estipula que -2.01 < *DioxEnxTot* <= 2.48 e o valor deste atributo no centróide é *DioxEnxTot* =0.487.

Observação (hipótese):

Neste *cluster*, ambas as médias da ***Massa Volúmica***, aparada a 5% ou não aparada, são iguais à da água (valor = 1), pontuação indicativa de baixo teor alcoólico (***Título Alcoométrico Volúmico Adquirido***). Para além disso, são identificados vinhos com um elevado ***Extrato Seco Total***, sendo mais espesso na boca, a presença de açúcares, cuja densidade é superior à da água. Assim sendo, estes valores influenciam directa e negativamente as variáveis de ***Aroma Qualidade***, ***Sabor Qualidade*** e ***Sabor Tipicidade***. Em resumo, o *cluster* é caracterisado por vinhos de qualidade inferior.

Este *cluster* é constituído maioritariamente por vinhos sem sub-região identificada (92%), sem presença significativa de nenhuma sub-região, conforme se constata pela seguinte tabela:

Sub-regiões	Nr amostras	Amostras %	c4	c4%
Amarante	501	3.49	34	2.16
Ave	471	3.28	31	1.97
Baiao	320	2.23	13	0.83
Basto	391	2.72	12	0.76
Cavado	431	3.00	7	0.44
Lima	465	3.24	3	0.19
Monção e Melgaço	1808	12.59	4	0.25
Paiva	155	1.08	3	0.19
Sousa	399	2.78	13	0.83
S/região	9414	65.58	1454	**92.38**
Totais	14355	100	1574	100

Cluster 5 – Vinhos com reduzido teor alcoólico

Características: 1126 elementos, que representa 7.8 % de dados.
A regra cobre 1175 amostras e estão correctamente cobertos 1105 amostras.

	Cluster 5	Características	InfoGain	Centróide	Centróide de Outros Clusters	Regra
1	AromaQualid	Sens.	0.636	4.981	6.182	
2	SaborQualid	Sens.	0.634	4.977	6.180	SaborQualid <= 5.5
3	SaborTipic	Sens.	0.557	5.061	6.237	
4	AromaTipic	Sens.	0.556	5.062	6.238	
5	TitAlcVolAdq	Fis.-Quim.	0.165	-0.650	-0.151	
6	TitAlcVolTot	Fis.-Quim.	0.162	-0.728	-0.144	
7	REGIAO	Identificação	0.131			
8	CASTA	Identificação	0.120			
9	MassVol	Fis.-Quim.	0.107	0.137	0.088	MassVol <= 1.69
10	TIPOI	Identificação	0.052			

Este agrupamento parece estar caracterizado por:

- Valor baixo de ***Sabor Qualidade***, *SaborQualid* (*InfoGain*=0.634). A regra estipula que *SaborQualid* <= 5.5 e o valor deste atributo no centróide é *SaborQualid*=4.977, extremamente baixo, quando comparado com o centróide dos outros *clusters*, 6.180.

- Valor de ***Massa Volúmica*** positivo mas baixo, *MassVol* (*InfoGain*=0.107). A regra estipula que *MassVol* <= 1.69 e o valor deste atributo no centróide é *MassVol* =0.137.

Observação (hipótese):
Cluster com maior ganho de informação nos atributos organolépticos. Ao nível das regras estabelecidas temos um valor de ***Sabor Qualidade*** baixo, cuja média, não aparada ou aparada a 5%, de 4.98 pontos, sendo classificado, na Escala de 0 a 10, como Medíocre. Ao nível físico-químico temos também um baixo valor de ***Massa Volúmica*** (0.99), cuja medida de referência

é a densidade da água (valor = 1). Deste modo, este *cluster* é caracterisado por vinhos de qualidade inferior.

Este *cluster* é constituído maioritariamente por vinhos sem sub-região identificada (95%), sem presença significativa de nenhuma sub-região, como se constata tabela seguinte:

Sub-regiões	Nr amostras	Amostras %	c5	c5%
Amarante	501	3.49	12	1.07
Ave	471	3.28	12	1.07
Baiao	320	2.23	2	0.18
Basto	391	2.72	11	0.98
Cavado	431	3.00	6	0.53
Lima	465	3.24	6	0.53
Monção e Melgaço	1808	12.59	7	0.62
Paiva	155	1.08	4	0.36
Sousa	399	2.78	2	0.18
S/região	9414	65.58	1064	94.49
Totais	14355	100	1126	100

Cluster 6 – Vinhos com acidez elevada

Características: 1411 elementos, que representa 9.8 % de dados.
A regra cobre 1459 amostras e estão correctamente cobertos 1340 amostras.

	Cluster 6	Características	InfoGain	Centróide	Centróide de Outros Clusters	Regra
1	AcidFix	Fis.-Quim.	0.616	1.627	-0.082	AcidFix > 0.51
2	AcidTot	Fis.-Quim.	0.599	1.609	-0.069	
3	ExtrNRed	Fis.-Quim.	0.284	1.155	-0.048	
4	SaborQualid	Sens.	0.169	5.758	6.122	SaborQualid <= 6.5
5	AromaQualid	Sens.	0.166	5.761	6.124	
6	AromaTipic	Sens.	0.165	5.824	6.181	
7	SaborTipic	Sens.	0.164	5.821	6.180	
8	RelAlcPeso_ExtNRed	Fis.-Quim.	0.123	-0.554	-0.076	
9	MassVol	Fis.-Quim.	0.093	0.124	0.089	
10	TitAlcVolAdq	Fis.-Quim.	0.083	-0.063	-0.209	
11	pH	Fis.-Quim.	0.058	-0.447	-0.042	
12	TitAlcVolTot	Fis.-Quim.	0.057	-0.151	-0.202	
20	REGIAO	Identificação	0.022			
26	TIPOI	Identificação	0.012			
27	CASTA	Identificação	0.012			

Este agrupamento parece estar caracterizado por:

- Valor elevado de ***Acidez Fixa***, AcidFix (*InfoGain*=0.616). A regra estipula que *Acidez Fixa* > 0.51 e o valor deste atributo no centróide é *Acidez Fixa* =1.627.

- Valor baixo de ***Sabor Qualidade***, *SaborQualid* (InfoGain=0.169). A regra estipula que *SaborQualid* <= 6.5 e o valor deste atributo no centróide é *SaborQualid*=5.758, abaixo do valor do centróide dos outros *clusters*, 6.122.

Observação (hipótese):

Cluster referenciado pela sua elevada acidez fixa, com médias de 7.58 pontos, advindo daí também o baixo ***Sabor Qualidade*** com média de 5.76, cuja Escala é de 0 a 10, e 5 o valor mais baixo de Suficiente.

Este *cluster* é constituído maioritariamente por vinhos sem sub-região identificada (64%), e temos também vinhos de Monção e Melgaço (6%), de Amarante (5%), do Ave (4%), do Cávado (4%), do Lima (4%) e do Sousa (4%), como pode ser constatado pela distribuição das amostras por regiões na seguinte tabela:

Sub-regiões	Nr amostras	Amostras %	c6	c6%
Amarante	501	3.49	67	4.75
Ave	471	3.28	62	4.39
Baiao	320	2.23	44	3.12
Basto	391	2.72	56	3.97
Cavado	431	3.00	61	4.32
Lima	465	3.24	58	4.11
Monção e Melgaço	1808	12.59	81	5.74
Paiva	155	1.08	25	1.77
Sousa	399	2.78	57	4.04
S/região	9414	65.58	900	63.78
Totais	14355	100	1411	100

Cluster 7 – Vinhos com Aroma e Sabor Qualidade baixo

Características: 242 elementos, que representa 1.7 % de dados.
A regra cobre 244 amostras e estão correctamente cobertos 244 amostras.

	Cluster 7	Características	InfoGain	Centróide	Centróide de Outros Clusters	Regra
1	AromaDefMarc	Sens.	1.000	1.0	1.999	
2	AromaQualid	Sens.	0.981	3.888	6.126	
3	SaborDefMarc	Sens.	0.981	1.004	1.999	
4	SaborQualid	Sens.	0.981	3.884	6.124	SaborQualid <= 4.5
5	AromaTipic	Sens.	0.615	5.0	6.165	
6	SaborTipic	Sens.	0.615	5.0	6.164	
13	CASTA	Identificação	0.046			
15	REGIAO	Identificação	0.042			
21	TIPOI	Identificação	0.016			

Este agrupamento parece estar caracterizado por:

- Valor baixo de ***Sabor Qualidade***, *SaborQualid* (*InfoGain*=0.981). A regra estipula que *SaborQualid* <= 4.5 e o valor deste atributo no centróide é *SaborQualid*=3.884 muito abaixo do centróide dos outros *clusters*, 6.124.

Observação (hipótese):

Cluster cujos valores apresentam ganhos e presença significativa ao nível dos atributos organolépticos. O item ***Sabor Qualidade*** apresenta uma média muito inferior à regra estabelecida. Deste modo estamos em presença de vinhos de uma específica qualidade.

Este *cluster* é constituído maioritariamente por vinhos sem sub-região identificada (80%), e temos também alguma presença de vinhos de Amarante (5%), como pode ser constatado pela distribuição das amostras por regiões na seguinte tabela:

Sub-regiões	Nr amostras	Amostras %	c7	c7%
Amarante	501	3.49	12	4.96
Ave	471	3.28	5	2.07
Baiao	320	2.23	4	1.65
Basto	391	2.72	8	3.31
Cavado	431	3.00	2	0.83
Lima	465	3.24	4	1.65
Monção e Melgaço	1808	12.59	8	**3.31**
Paiva	155	1.08	3	1.24
Sousa	399	2.78	2	0.83
S/região	9414	65.58	194	**80.17**
Totais	14355	100	242	100

Cluster 8 – Vinhos com elevada Limpidez

Características: 330 elementos, que representa 2.3 % de dados.
A regra cobre 334 amostras e estão correctamente cobertos 330 amostras.

	Cluster 8	Características	InfoGain	Centróide	Centróide de Outros Clusters	Regra
1	AspLimpid	Sens.	0.953	2.524	1.019	AspLimpid > 1.46
2	AromaQualid	Sens.	0.138	5.409	6.104	
3	SaborQualid	Sens.	0.137	5.409	6.102	
4	SaborTipic	Sens.	0.135	5.521	6.160	
5	AromaTipic	Sens.	0.135	5.521	6.160	
6	AcucarRed	Fis.-Quim.	0.109	-0.499	0.008	
7	AspCor	Sens.	0.056	0.634	0.113	
8	AcidAsc	Fis.-Quim.	0.055	0.006	-0.009	
9	ExtrSecTot	Fis.-Quim.	0.054	-0.452	0.038	
10	RelAlcPeso_ExtNRed	Fis.-Quim.	0.054	-0.321	-0.099	
19	REGIAO	Identificação	0.026			
23	CASTA	Identificação	0.022			
24	TIPOI	Identificação	0.010			

Este agrupamento parece estar caracterizado por:

- Valor positivo muito elevado de ***Aspecto Limpidez***, *AspLimpid* (*InfoGain*=0.953). A regra estabelece que *AspLimpid* > 1.46 e o valor deste atributo no centróide é *AspLimpid* =2.524, muito superior à dos centróides dos outros *clusters*, 1.019.

Observação (hipótese):

Cluster cujos valores apresentam ganhos e presença significativa ao nível do atributo organoléptico ***Aspecto Limpidez***, sendo, provavelmente, composto por vinhos de qualidade superior.

Este *cluster* é constituído maioritariamente por vinhos sem sub-região identificada (71%), e temos também a presença de vinhos do Ave (5%) e de Monção e Melgaço (5%), como pode ser constatado pela distribuição das amostras por regiões na seguinte tabela:

Sub-regiões	Nr amostras	Amostras %	c8	c8%
Amarante	501	3.49	11	3.33
Ave	471	3.28	17	5.15
Baiao	320	2.23	11	3.33
Basto	391	2.72	9	2.73
Cavado	431	3.00	8	2.42
Lima	465	3.24	12	3.64
Monção e Melgaço	1808	12.59	18	**5.45**
Paiva	155	1.08	4	1.21
Sousa	399	2.78	5	**1.52**
S/região	9414	65.58	235	71.21
Totais	14355	100	330	100

Cluster 11 – Vinhos com Cloretos elevados

Características: 306 elementos, que representa 2.1 % de dados.
A regra cobre 301 amostras e estão correctamente cobertos 299 amostras.

	Cluster 11	Características	InfoGain	Centróide	Centróide de Outros Clusters	Regra
1	Cloret	Fis.-Quim.	0.916	5.478	-0.041	Cloret > 1.97
2	AcidCitric	Fis.-Quim.	0.314	2.098	-0.037	
3	TitAlcVolAdq	Fis.-Quim.	0.255	-1.006	-0.115	
4	TitAlcVolTot	Fis.-Quim.	0.220	-0.947	-0.122	
5	SaborTipic	Sens.	0.112	5.667	6.155	
6	AromaTipic	Sens.	0.111	5.669	6.156	
7	AromaQualid	Sens.	0.106	5.641	6.098	
8	SaborQualid	Sens.	0.105	5.641	6.096	
9	REGIAO	Identificação	0.102			
10	MassVol	Fis.-Quim.	0.099	0.366	0.065	
11	pH	Fis.-Quim.	0.072	-0.578	-0.029	
12	CASTA	Identificação	0.071			
13	RelAlcPeso_ExtNRed	Fis.-Quim.	0.070	-0.269	-0.104	
14	TIPOI	Identificação	0.053			

Este agrupamento parece estar caracterizado por:

- Valor elevado de ***Cloretos***, *Cloret* (*InfoGain*=0.916). A regra estipula que *Cloret* > 1.97 e o valor deste atributo no centróide é *Cloret =5.478*.

Observação (hipótese):

Cluster cujo item com valores mais expressivos são os ***Cloretos***, cuja presença pode ser indicativa de vinhos provenientes de solos salgados (Alpuim, 1997). Acresce que os centróides das características organolépticas indicam valores baixos pelo que este *cluster* é caracterisados por vinhos de qualidade inferior.

Este *cluster* é constituído maioritariamente por vinhos sem sub-região identificada (92%), sem presença significativa de nenhuma sub-região, como pode ser constatado pela distribuição das amostras por regiões na seguinte tabela:

Sub-regiões	Nr amostras	Amostras %	c11	c11%
Amarante	501	3.49	7	2.29
Ave	471	3.28	1	0.33
Baiao	320	2.23	1	0.33
Basto	391	2.72	2	0.65
Cavado	431	3.00	3	0.98
Lima	465	3.24	1	0.33
Monção e Melgaço	1808	12.59	7	2.29
Paiva	155	1.08	0	0.00
Sousa	399	2.78	1	0.33
S/região	9414	65.58	283	**92.48**
Totais	14355	100	306	100

Anexo B1

Características usadas nos dados

Características gerais (8 atributos)

Ano Prova – ano em que o vinho foi dado à prova

Regiao – região vinícola portuguesa

Categoria - engloba todas as denominações que uma colheita pode ter, i.e, premium, standard, roses adamados, entre outros.

Cor – tipo de vinho: branco, tinto e rose

Data Colheita – data em que o vinho foi colhido

CodVinho – código da marca do vinho

Data Prova – data específica da prova

Temp Vinhos (°C) – temperatura do vinho à data da prova

Características físico-químicas (ensaios analíticos) (30 atributos)

Acidez volátil (gL ac.acet.) - Componente presente no vinho que, em dose elevada, origina o aroma a vinagre. Em excesso é o resultado da falta de cuidados durante a vinificação. Nos vinhos velhos é habitual um toque mais acentuado de acidez volátil, aos quais se dá a designação de "vinagrinho". (Sogrape, 2016)

pH - Potencial de hidrogénio. Índice de acidez ou de alcalinidade de um vinho. Calcula-se pelo co-logaritmo de concentração em iões hidrogénio. Quando o valor em pH é inferior a 7, o líquido é ácido. Os solos da vinha distinguem-se também pelo seu pH maior ou menor. O pH de um mosto varia entre 2,8 a 3,8; o pH de um vinho, varia de 3 a 4.(Sogrape, 2016)

Alcool Inf (vv) – grau de álcool.

SO2 livre (mgL) – presença de Dióxido de Enxofre Livre.

SO2 total (mgL) – presença de Dióxido de Enxofre Total.

SO2 molec (mgL) – presença de Dióxido de Enxofre molecular.

Densidade (gL) - Conjunto de sensações de leveza, suavidade, elegância e espacialidade de um vinho. A densidade de um vinho pode ser descrita como Leve até Sólida, passando por Aveludada, Glicerinada, Redondo, Sedoso, Elegante, Untuoso, Amplo, Pleno, Cheio, Carnudo ou Robusto (Coutinho, 2016).

BE - grau Baumé, medida densimétrica que corresponde à quantidade de açúcar existente no vinho – aplicável apenas a Vinhos do Porto (Sogrape, 2016).

Acidez total (gL ac.tart.) - ou acidez de titulação, soma dos ácidos tituláveis, quando se eleva o vinho a pH7, pela adição de uma solução alcalina titulável (o CO2 e o SO2 total não estão compreendidos neste valor) (Sogrape, 2016).

ABS 420 (nm) – Densidade óptica a 420 nm.

ABS 520 (nm) – Densidade óptica a 520 nm.

Anexo B1 (cont.)

ABS 620 (nm) – Densidade óptica a 620 nm.

Intensidade - Adjetivo utilizado na prova para distinguir certas qualidades: cor, aroma ou sabor profundos. Implica certa complexidade na composição do vinho e uma duração sustentada das impressões organolépticas.

Tonalidade - Adjetivo utilizado no primeiro exame visual do vinho para distinguir certas qualidades: idade, corpo, viscosidade, teor alcoólico, tipo de uva que o compõem, entre outras.

IPT (índice de polifenóis totais) – Taninos, antocianos e outras matérias corantes de gosto amargo, que contribuem também para a estrutura dos vinhos (Sogrape, 2016).

Extracto seco (gL) - Matérias secas totais; conjunto de todas as substâncias que, em determinadas condições de temperatura e pressão, não se volatilizam. Exprime-se em gramas por litro. Um vinho com pouco extracto é leve; com muito é espesso na boca (Sogrape, 2016).

Extracto nao redutor (gL) - extracto seco total diminuído dos açúcares totais (I.V.D.P., 2016).

Substâncias red - Substâncias redutoras

Glucose+Fructose - açúcares (expresso em *glucose+frutose*) A glucose e a frutose são doseadas individualmente por um método enzimático, com vista unicamente ao cálculo da razão glucose/frutose, ou seja, ao cálculo do tempo necessário para a acção das enzimas (JOUE, 2010)

Fenóis Voláteis (µgL) – (Fenol - Composto derivado de um núcleo benzénico por reutilização dos átomos de hidrogénio por um ou vários grupos hidróxilos. Os compostos fenólicos (ácidos fenóis, álcoois fenóis, aldeídos fenóis, antocianinas, flavonas, taninos) têm um papel determinante na cor, na adstringência e na composição aromática de um vinho (Sogrape, 2016).

Antocianas (mgL) - Compostos fenólicos ou pigmentos de cor vermelho-arroxeado, contidos na película das uvas tintas. A proporção de antocianas presente no vinho é de 0,25 a 0,5 gramas por litro. Estes pigmentos são sensíveis à oxidação. Por este motivo, a cor dos vinhos evolui do violáceo para o vermelho, do rubi para o ocre-escuro e dos tons da telha para o queimado-alaranjado. As antocianas exercem um efeito saudável sobre as artérias, ao aumentar a taxa de colesterol bom (lipoproteínas de alta densidade). Têm também uma influência saudável pelo seu efeito face aos radicais livres (Infovini, 2016).

Taninos (gL) - Conjunto dos compostos fenólicos de um vinho, responsáveis pela sua cor, o seu aroma, a sua estrutura e muitas outras virtudes. Substância orgânica de sabor adstringente, contida nas películas e nas grainhas da uva (Sogrape, 2016).

Turbidez (NTU) - Medição da turbidez (NTU) - medição da turvação provocada pela difusão (D.R.A.P.C., 2016).

Fe (mgL) – Sabor metálico característico de certos solos. Pode ser um defeito grave quando se deve a uma casse férrica, no caso de os vinhos terem entrado em contacto com objectos de ferro (Infovini, 2016).

Anexo B1 (cont.)

Cu (mgL) – Cobre - Odor desagradável de um vinho alterado e estragado pela presença de cobre (casse cúprica). O excesso de cobre (mais de 1 mg/l) detecta-se imediatamente nos vinhos brancos devido à sua cor parda. (CVRVV, 2016)

Ca (mgL) - elemento químico Cálcio

Acetato de Etilo (mgL) - esterificação do ácido acético por bactérias acéticas (Carvalheira, 2011)

4-etil-fenol – fenol utilizado em vinho e cerveja, é produzido pela deterioração da levedura Brettanomyces (wikipedia).

4-etil-gaiacol – fenol produzido juntamente como o 4-etilfenol (4-EP, do inglês 4-ethylphenol) em vinho e cerveja pela deterioração da levedura Brettanomyces (wikipedia).

Classificações dos provadores *(16 atributos)* - P112; P284; P384; P441; P444; P467; P555; P735; P736; P800; P828; P888; P993; P911; P924 e P938.

Anexo B2

	Variável	Descrição	Codificação	Tipo de variável
Dados identificadores do vinho	AnoProva	Ano em que vinho foi dado à prova	AnoProva	Discreta
	Categoria	Denominações de uma colheita	Alvarinho	Categórica
			Premiukm	
			Standard	
	CodVinho	Código identificativo da marca e lote do vinho		Discreta
	Cor	Aspecto cromático do vinho	Branco	Categórica
	DataColheita	Data em que é efectuada a vindima		
	DataProva	Data específica da prova	DataProva	
	Regiao	Região vinícola demarcada	Alentejo	Categórica
			Dão	
			Douro	
			Vinhos verdes	
Características Físico-químicas	ABS420nm	Densidade óptica	0.034 a 0.202	Contínua
	AcidTot	Acidez Total	3.4 a 11.7	Contínua
	AcidVolat	Acidez volátil	0.18 a 0.52	Contínua
	Alcool_Vol%	Grau alcoolémico	8.13 a 14.49	Contínua
	Densidade	Densidade	0.988 a 1.01	Contínua
	ExtrSeco	Extracto Seco	0.1 a 52.7	Contínua
	pH	pH	2.81 a 3.58	Contínua
	SO2livre	Dióxido de enxofre livre	7 a 101	Discreta
	SO2total	Dióxido de enxofre total	49 a 218	Discreta
	SubstRed	Substancias Redutoras	0 a 16.7	Contínua
Painel de Provadores	Defeit	Defeituoso	Sim/Não	Categórica
	Nr Provad	Nr Provadores	4 a 11	Discreta
	P112	Provador 112	6 a 18	Contínua
	P284	Provador 284	9 a 18	Contínua
	P384	Provador 384	9 a 17.5	Contínua
	P441	Provador 441	0 a 19	Contínua
	P444	Provador 444	10 a 19	Contínua
	P467	Provador 467	7 a 17.5	Contínua
	P555	Provador 555	10 a 17	Contínua
	P735	Provador 735	0 a 18	Contínua
	P736	Provador 736	0 a 18	Contínua
	P800	Provador 800	9 a 17	Contínua
	P828	Provador 828	8.8 a 17	Contínua
	P911	Provador 911	0 a 18	Discreta
	P924	Provador 924	0 a 18	Contínua
	Rank	Ranking	1 a 14.3	Discreta
Parâmetros descritivos da Nota	CoefVar	Coeficiente Variacao	-0.3 a 1.23	Contínua
	DesvPadr	DesvioPadrao	-0.48 a 6.01	Contínua
	Iqant	IQantigo	0.1 a 2.85	Contínua
	Iqatual	IQatual	0.22 a 8.56	Contínua
	MediaAbs	MediaAbsoluta	4.89 a 17.38	Contínua
	MediafetadaPreco	Media afetada pelo preco	0 a 19.03	Contínua
	NotaAmplitude	NotaAmplitude	0 a 14	Contínua
	NotaInterquantil	NotaInterquantil	0 a 10	Contínua
	NotaMax	NotaMax	10 a 19	Contínua
	NotaMediana	NotaMediana	0 a 17.8	Contínua
	NotaMin	NotaMin	0 a 16	Contínua
	NotaQ1	NotaQ1	0 a 17.1	Contínua
	NotaQ3	NotaQ3	10 a 18	Contínua
	Preco	Preco	1.35 a 39.5	Contínua

Anexo B3

Test t dos Provadores

Provadores	t	df	p-value (2-tailed)	Limite Inferior	Limite Superior
P441 - MediaO	1.483	283	0.139	-0.042	0.300
P467 - MediaO	2.880	237	0.004	0.083	0.044
P284 - MediaO	1.450	218	0.148	-0.049	0.323
P924 - MediaO	-4.100	212	0.000	-0.642	-0.225
P911 - MediaO	-4.416	208	0.000	-0.510	-0.195
P444 - MediaO	3.221	208	0.002	0.115	0.476
P736 - MediaO	-1.528	205	0.128	-0.450	0.057
P112 - MediaO	-1.805	181	0.073	-0.504	0.023
P800 - MediaO	2.644	179	0.009	0.056	0.389
P555 - MediaO	1.464	92	0.147	-0.080	0.529
P384 - MediaO	0.236	82	0.814	-0.302	0.383
P828 - MediaO	-1.373	75	0.174	-0.625	0.115
P735 - MediaO	-0.223	48	0.825	-0.561	0.449
P888 - MediaO	2.073	25	0.049	0.004	1.241
P938 - MediaO	0.260	9	0.801	-0.463	0.583

Analisando os limites inferiores e superiores da tabela acima apresentada também se pode verificar os intervalos de atribuição de notas de cada provador nas diversas provas, verificando, caso os dois limites sejam negativos, que o provador atribuiu sempre notas abaixo da média, sendo esta considerada como zero.

Participações e ausências de cada provador nas provas de vinhos

	Casos			
	Válido		Ausente	
	N	Percentagem	N	Percentagem
P888	25	7,5%	310	92,5%
P444	209	62,4%	126	37,6%
P467	238	71,0%	97	29,0%
P800	180	53,7%	155	46,3%
P555	93	27,8%	242	72,2%
P284	219	65,4%	116	34,6%
P441	284	84,8%	51	15,2%
P938	10	3,0%	325	97,0%
P384	83	24,8%	252	75,2%
P735	49	14,6%	286	85,4%
P112	182	54,3%	153	45,7%
P736	206	61,5%	129	38,5%
P828	76	22,7%	259	77,3%
P911	209	62,4%	126	37,6%
P924	213	63,6%	122	36,4%

Anexo B4

M5P - M5 pruned model tree: (using smoothed linear models)

Amostra <= 201.5 :
| AnoProva <= 2008.5 :
| | CumgL <= 0.075 : LM1 (165/85.649%)
| | CumgL > 0.075 :
| | | SO2molecmgL <= 1.089 : LM2 (81/117.506%)
| | | SO2molecmgL > 1.089 : LM3 (130/92.915%)
| AnoProva > 2008.5 :
| | Substanciasred <= 5.05 :
| | | SO2livremgL <= 27.5 :
| | | | Avaliador <= 869.5 : LM4 (229/69.179%)
| | | | Avaliador > 869.5 : LM5 (65/70.15%)
| | | SO2livremgL > 27.5 :
| | | | DensidadegL <= 0.992 : LM6 (375/67.921%)
| | | | DensidadegL > 0.992 :
| | | | | pH <= 3.2 :
| | | | | | SO2livremgL <= 33.5 :
| | | | | | | Avaliador <= 362.5 : LM7 (7/90.723%)
| | | | | | | Avaliador > 362.5 : LM8 (26/46.192%)
| | | | | | SO2livremgL > 33.5 : LM9 (35/66.518%)
| | | | | pH > 3.2 : LM10 (28/69.438%)
| | Substanciasred > 5.05 :
| | | CumgL <= 0.092 :
| | | | Amostra <= 150.5 :
| | | | | SO2totalmgL <= 124.5 : LM11 (18/61.97%)
| | | | | SO2totalmgL > 124.5 : LM12 (24/73.289%)
| | | | Amostra > 150.5 : LM13 (47/68.57%)
| | | CumgL > 0.092 :
| | | | SO2totalmgL <= 117.5 : LM14 (70/82.152%)
| | | | SO2totalmgL > 117.5 :
| | | | | Avaliador <= 768 : LM15 (48/79.808%)
| | | | | Avaliador > 768 : LM16 (17/35.801%)
Amostra > 201.5 :
| AcidezvolatilgLacacet <= 0.415 :
| | Substanciasred <= 2.95 : LM17 (341/79.721%)
| | Substanciasred > 2.95 :
| | | Substanciasred <= 4.25 :
| | | | SO2totalmgL <= 96 : LM18 (34/106.689%)
| | | | SO2totalmgL > 96 :
| | | | | ABS420nm <= 0.057 : LM19 (31/60.073%)
| | | | | ABS420nm > 0.057 : LM20 (91/103.71%)
| | | Substanciasred > 4.25 : LM21 (164/66.912%)
| AcidezvolatilgLacacet > 0.415 :
| | SO2molecmgL <= 0.917 :
| | | SO2totalmgL <= 105 :
| | | | SO2molecmgL <= 0.586 :
| | | | | Amostra <= 253.5 : LM22 (17/86.255%)
| | | | | Amostra > 253.5 : LM23 (6/83.036%)
| | | | SO2molecmgL > 0.586 : LM24 (33/103.367%)
| | | SO2totalmgL > 105 : LM25 (42/60.608%)
| | SO2molecmgL > 0.917 : LM26 (144/62.808%)

Anexo B4 (cont.)

LM num: 1
Nota =
0.0003 * Amostra
- 0 * Avaliador
+ 0.0034 * AnoProva
+ 0.0078 * CodVinho
+ 0.4816 * AcidezvolatilgLacacet
+ 8.6002 * pH
+ 0.7832 * AlcoolInfvv
+ 0.0066 * SO2livremgL
+ 0 * SO2totalmgL
- 0.0659 * SO2molecmgL
+ 39.8535 * DensidadegL
+ 1.7217 * AcideztotalgLactart
- 0.551 * ABS420nm
+ 0.0087 * IPT
- 0.0013 * ExtractosecogL
- 0.0079 * ExtractonaoredgL
- 0.0424 * Substanciasred
- 0.0466 * TurbidezNTU
+ 0.7183 * FemgL
- 0.5595 * CumgL
- 81.7808

LM num: 2
Nota =
0.0011 * Amostra
- 0 * Avaliador
+ 0.0034 * AnoProva
+ 0.0007 * CodVinho
- 0.5364 * AcidezvolatilgLacacet
- 0.509 * pH
- 0.0051 * AlcoolInfvv
+ 0.0054 * SO2livremgL
+ 0 * SO2totalmgL
- 0.0525 * SO2molecmgL
+ 124.6617 * DensidadegL
+ 0.0478 * AcideztotalgLactart
- 33.8779 * ABS420nm
+ 0.0366 * IPT
- 0.0013 * ExtractosecogL
- 0.0383 * ExtractonaoredgL
- 0.117 * Substanciasred
- 0.0371 * TurbidezNTU
- 0.7042 * FemgL
- 0.4836 * CumgL
- 111.3444

LM num: 3
Nota =
0.0008 * Amostra
- 0 * Avaliador
+ 0.9706 * AnoProva
- 0.0081 * CodVinho
- 0.2187 * AcidezvolatilgLacacet
- 0.3407 * pH
- 0.0051 * AlcoolInfvv
+ 0.0054 * SO2livremgL
+ 0 * SO2totalmgL
- 0.79 * SO2molecmgL
+ 93.8585 * DensidadegL
+ 0.0478 * AcideztotalgLactart
- 0.551 * ABS420nm
+ 0.0265 * IPT
- 0.0013 * ExtractosecogL
- 0.0275 * ExtractonaoredgL
- 0.0041 * Substanciasred
- 0.0371 * TurbidezNTU
- 0.1092 * FemgL
- 0.4836 * CumgL
- 2025.5242

LM num: 4
Nota =
0.0001 * Amostra
- 0.0001 * Avaliador
+ 0.0013 * AnoProva
- 0 * CodVinho
+ 0.0577 * AcidezvolatilgLacacet
- 1.431 * pH
+ 0.0452 * AlcoolInfvv
+ 0.0856 * SO2livremgL
- 0.0135 * SO2totalmgL
- 0.0058 * SO2molecmgL
+ 5.8404 * DensidadegL
+ 0.0105 * AcideztotalgLactart
- 21.4928 * ABS420nm
+ 0.0011 * IPT
+ 0.0029 * ExtractosecogL
+ 0.0001 * ExtractonaoredgL
- 0.0015 * Substanciasred
+ 0.0093 * FemgL
- 0.2627 * CumgL
+ 10.3565

LM num: 5
Nota =
-0.0058 * Amostra
- 0.0002 * Avaliador
+ 0.0013 * AnoProva
- 0 * CodVinho
+ 0.0577 * AcidezvolatilgLacacet
- 0.2686 * pH
+ 0.0811 * AlcoolInfvv
+ 0.0908 * SO2livremgL
- 0.0207 * SO2totalmgL
- 0.0058 * SO2molecmgL
+ 5.8404 * DensidadegL
+ 0.0321 * AcideztotalgLactart
- 4.0417 * ABS420nm
+ 0.0531 * IPT
+ 0.071 * ExtractosecogL
+ 0.0001 * ExtractonaoredgL
- 0.0015 * Substanciasred
- 0.6547 * FemgL
- 0.2627 * CumgL
+ 3.994

LM num: 6
Nota =
0.0001 * Amostra
- 0 * Avaliador
+ 0.0013 * AnoProva
- 0 * CodVinho
+ 0.0577 * AcidezvolatilgLacacet
+ 1.359 * pH
+ 0.0233 * AlcoolInfvv
- 0.011 * SO2livremgL
+ 0 * SO2totalmgL
- 0.0058 * SO2molecmgL
+ 1.5678 * DensidadegL
- 0.5602 * ABS420nm
- 0.0009 * IPT
- 0 * ExtractosecogL
+ 0.0013 * ExtractonaoredgL
- 0.0015 * Substanciasred
+ 0.0073 * FemgL
- 0.1136 * CumgL
+ 5.7007

Anexo B4 (cont.)

LM num: 7
Nota =
-0.0223 * Amostra
+ 0.0008 * Avaliador
+ 0.0013 * AnoProva
+ 0.0038 * CodVinho
+ 0.0577 * AcidezvolatilgLacacet
+ 6.2178 * pH
- 0.5631 * AlcoolInfvv
- 0.0178 * SO2livremgL
+ 0 * SO2totalmgL
- 0.0058 * SO2molecmgL
- 9.1712 * DensidadegL
- 0.5602 * ABS420nm
- 0.0009 * IPT
- 0 * ExtractosecogL
+ 0.0042 * ExtractonaoredgL
- 0.0015 * Substanciasred
+ 0.0073 * FemgL
+ 0.1724 * CumgL
+ 9.0495

LM num: 8
Nota =
-0.0185 * Amostra
+ 0.0004 * Avaliador
+ 0.0013 * AnoProva
+ 0.0038 * CodVinho
+ 0.0577 * AcidezvolatilgLacacet
+ 3.9119 * pH
- 0.2756 * AlcoolInfvv
+ 0.0686 * SO2livremgL
+ 0 * SO2totalmgL
- 0.0058 * SO2molecmgL
- 9.1712 * DensidadegL
- 0.5602 * ABS420nm
- 0.0009 * IPT
- 0 * ExtractosecogL
+ 0.0042 * ExtractonaoredgL
- 0.0015 * Substanciasred
+ 0.0073 * FemgL
+ 0.1724 * CumgL
+ 10.4068

LM num: 9
Nota =
-0.0135 * Amostra
- 0.0015 * Avaliador
+ 0.0013 * AnoProva
+ 0.0036 * CodVinho
+ 0.0577 * AcidezvolatilgLacacet
+ 1.242 * pH
- 0.121 * AlcoolInfvv
- 0.0178 * SO2livremgL
+ 0 * SO2totalmgL
- 0.0058 * SO2molecmgL
- 9.1712 * DensidadegL
- 0.5602 * ABS420nm
- 0.0009 * IPT
- 0 * ExtractosecogL
+ 0.0042 * ExtractonaoredgL
- 0.0015 * Substanciasred
+ 0.0073 * FemgL
+ 0.1724 * CumgL
+ 19.4947

LM num: 10
Nota =
-0.0082 * Amostra
- 0 * Avaliador
+ 0.0013 * AnoProva
- 0 * CodVinho
+ 0.0577 * AcidezvolatilgLacacet
+ 2.0803 * pH
- 0.0428 * AlcoolInfvv
- 0.0329 * SO2livremgL
+ 0 * SO2totalmgL
- 0.0058 * SO2molecmgL
- 9.1712 * DensidadegL
- 0.5602 * ABS420nm
- 0.0009 * IPT
- 0 * ExtractosecogL
+ 0.0042 * ExtractonaoredgL
- 0.0015 * Substanciasred
+ 0.0073 * FemgL
+ 0.1724 * CumgL
+ 16.5705

LM num: 11
Nota =
-0.0077 * Amostra
- 0.0001 * Avaliador
+ 0.0013 * AnoProva
- 0.0036 * CodVinho
+ 0.0577 * AcidezvolatilgLacacet
+ 0.2728 * pH
+ 0.0307 * AlcoolInfvv
+ 0.0017 * SO2livremgL
- 0.0123 * SO2totalmgL
- 0.019 * SO2molecmgL
+ 8.7783 * DensidadegL
- 6.7009 * ABS420nm
- 0.0018 * IPT
- 0.0006 * ExtractosecogL
- 0.0141 * ExtractonaoredgL
- 0.0015 * Substanciasred
+ 0.0123 * FemgL
- 0.8515 * CumgL
+ 4.9127

LM num: 12
Nota =
0.0016 * Amostra
- 0.0001 * Avaliador
+ 0.0013 * AnoProva
- 0.0036 * CodVinho
+ 0.0577 * AcidezvolatilgLacacet
+ 0.2728 * pH
+ 0.0307 * AlcoolInfvv
+ 0.0017 * SO2livremgL
- 0.011 * SO2totalmgL
- 0.019 * SO2molecmgL
+ 8.7783 * DensidadegL
- 6.7009 * ABS420nm
- 0.0018 * IPT
- 0.0006 * ExtractosecogL
- 0.0141 * ExtractonaoredgL
- 0.0015 * Substanciasred
+ 0.0123 * FemgL
- 0.8515 * CumgL
+ 3.254

Anexo B4 (cont.)

LM num: 13
Nota =

0.0057 * Amostra
- 0.0001 * Avaliador
+ 0.0013 * AnoProva
- 0.0033 * CodVinho
+ 0.0577 * AcidezvolatilgLacacet
+ 0.2499 * pH
+ 0.0307 * AlcoolInfvv
+ 0.0017 * SO2livremgL
- 0.003 * SO2totalmgL
- 0.019 * SO2molecmgL
+ 8.7783 * DensidadegL
- 6.3914 * ABS420nm
- 0.0018 * IPT
- 0.0006 * ExtractosecogL
- 0.0141 * ExtractonaoredgL
- 0.0015 * Substanciasred
+ 0.0123 * FemgL
- 0.8515 * CumgL
+ 2.7137

LM num: 14
Nota =

-0.0012 * Amostra
- 0.0002 * Avaliador
+ 0.0013 * AnoProva
+ 0.0006 * CodVinho
- 20.2056 * AcidezvolatilgLacacet
- 0.0108 * pH
+ 0.0307 * AlcoolInfvv
+ 0.0017 * SO2livremgL
+ 0.0523 * SO2totalmgL
- 0.019 * SO2molecmgL
+ 8.7783 * DensidadegL
- 3.6262 * ABS420nm
- 0.0018 * IPT
- 0.0006 * ExtractosecogL
- 0.0377 * ExtractonaoredgL
- 0.0015 * Substanciasred
+ 0.0123 * FemgL
- 0.6935 * CumgL
+ 2.3478

LM num: 15
Nota =

0.0042 * Amostra
- 0.0021 * Avaliador
+ 0.0013 * AnoProva
+ 0.0014 * CodVinho
- 1.5436 * AcidezvolatilgLacacet
- 0.0108 * pH
+ 0.0307 * AlcoolInfvv
+ 0.0017 * SO2livremgL
+ 0.0077 * SO2totalmgL
- 0.019 * SO2molecmgL
+ 8.7783 * DensidadegL
- 3.7135 * ABS420nm
- 0.0018 * IPT
- 0.0006 * ExtractosecogL
- 0.0699 * ExtractonaoredgL
- 0.0015 * Substanciasred
+ 0.0123 * FemgL
- 0.6935 * CumgL
+ 2.8968

LM num: 16
Nota =

0.0057 * Amostra
- 0.0006 * Avaliador
+ 0.0013 * AnoProva
+ 0.0022 * CodVinho
- 1.5436 * AcidezvolatilgLacacet
- 0.0108 * pH
+ 0.0307 * AlcoolInfvv
+ 0.0017 * SO2livremgL
+ 0.0077 * SO2totalmgL
- 0.019 * SO2molecmgL
+ 8.7783 * DensidadegL
+ 5.2909 * ABS420nm
- 0.0018 * IPT
- 0.0006 * ExtractosecogL
- 0.0994 * ExtractonaoredgL
- 0.0015 * Substanciasred
+ 0.0123 * FemgL
- 0.6935 * CumgL
+ 2.1564

LM num: 17
Nota =

-0.0122 * Amostra
- 0.0006 * Avaliador
+ 0.0047 * AnoProva
+ 0 * CodVinho
+ 5.3253 * AcidezvolatilgLacacet
- 0.0163 * pH
+ 0.0066 * AlcoolInfvv
+ 0.0009 * SO2livremgL
+ 0.0001 * SO2totalmgL
+ 0.0061 * SO2molecmgL
- 246.3365 * DensidadegL
+ 0.1683 * AcideztotalgLactart
- 8.0672 * ABS420nm
- 0.0002 * IPT
+ 0.157 * ExtractosecogL
- 0.0004 * ExtractonaoredgL
- 0.0011 * Substanciasred
+ 0.0138 * CumgL
+ 246.6827

LM num: 18
Nota =

-0.0239 * Amostra
- 0.0003 * Avaliador
+ 0.005 * AnoProva
+ 0.0002 * CodVinho
+ 0.5322 * AcidezvolatilgLacacet
+ 0.8299 * pH
- 0.2551 * AlcoolInfvv
+ 0.0338 * SO2livremgL
+ 0.0001 * SO2totalmgL
+ 0.0061 * SO2molecmgL
- 229.9364 * DensidadegL
+ 0.0114 * AcideztotalgLactart
+ 2.15 * ABS420nm
- 0.0002 * IPT
+ 0.0057 * ExtractosecogL
- 0.0004 * ExtractonaoredgL
+ 0.0249 * Substanciasred
+ 0.0138 * CumgL
+ 235.9762

Anexo B4 (cont.)

LM num: 19
Nota =
-0.0023 * Amostra
- 0.0004 * Avaliador
+ 0.005 * AnoProva
+ 0.0002 * CodVinho
- 7.0565 * AcidezvolatilgLacacet
+ 0.2864 * pH
+ 0.0228 * AlcoolInfvv
+ 0.0145 * SO2livremgL
+ 0.0001 * SO2totalmgL
+ 0.0061 * SO2molecmgL
- 105.8391 * DensidadegL
+ 0.0114 * AcideztotalgLactart
- 2.5618 * ABS420nm
- 0.0002 * IPT
+ 0.0057 * ExtractosecogL
- 0.0004 * ExtractonaoredgL
+ 0.0249 * Substanciasred
+ 0.0138 * CumgL
+ 110.5598

LM num: 20
Nota =
-0.0031 * Amostra
- 0.0011 * Avaliador
+ 0.005 * AnoProva
+ 0.0002 * CodVinho
- 0.3141 * AcidezvolatilgLacacet
+ 0.2864 * pH
- 0.0466 * AlcoolInfvv
+ 0.0145 * SO2livremgL
+ 0.0001 * SO2totalmgL
+ 0.0061 * SO2molecmgL
- 338.5248 * DensidadegL
+ 0.0114 * AcideztotalgLactart
- 0.7581 * ABS420nm
- 0.0002 * IPT
+ 0.0057 * ExtractosecogL
- 0.0004 * ExtractonaoredgL
+ 0.0249 * Substanciasred
+ 0.0138 * CumgL
+ 340.2016

LM num: 21
Nota =
0.01 * Amostra
- 0.0001 * Avaliador
- 0.1175 * AnoProva
+ 0.0024 * CodVinho
+ 7.7158 * AcidezvolatilgLacacet
- 1.7115 * pH
- 0.0126 * AlcoolInfvv
+ 0.0037 * SO2livremgL
+ 0.0001 * SO2totalmgL
+ 0.0061 * SO2molecmgL
- 35.7667 * DensidadegL
+ 0.0114 * AcideztotalgLactart
- 0.2246 * ABS420nm
- 0.0002 * IPT
+ 0.0057 * ExtractosecogL
- 0.0004 * ExtractonaoredgL
+ 0.0237 * Substanciasred
+ 0.0138 * CumgL
+ 285.829

LM num: 22
Nota =
-0.0075 * Amostra
- 0 * Avaliador
+ 0.0079 * CodVinho
+ 0.3822 * AcidezvolatilgLacacet
- 0.0163 * pH
+ 0.0971 * AlcoolInfvv
+ 0.0006 * SO2livremgL
+ 0.0014 * SO2totalmgL
+ 2.1371 * SO2molecmgL
- 14.198 * DensidadegL
+ 0.044 * AcideztotalgLactart
- 0.3813 * ABS420nm
- 0.0002 * IPT
+ 0.0054 * ExtractosecogL
- 0.0004 * ExtractonaoredgL
- 0.0011 * Substanciasred
+ 0.0471 * FemgL
+ 0.1077 * CumgL
+ 26.1398

LM num: 23
Nota =
-0.0075 * Amostra
- 0 * Avaliador
+ 0.0079 * CodVinho
+ 0.3822 * AcidezvolatilgLacacet
- 0.0163 * pH
+ 0.0971 * AlcoolInfvv
+ 0.0006 * SO2livremgL
+ 0.0014 * SO2totalmgL
+ 2.1371 * SO2molecmgL
- 14.198 * DensidadegL
+ 0.044 * AcideztotalgLactart
- 0.3813 * ABS420nm
- 0.0002 * IPT
+ 0.0054 * ExtractosecogL
- 0.0004 * ExtractonaoredgL
- 0.0011 * Substanciasred
+ 0.0471 * FemgL
+ 0.1077 * CumgL
+ 25.7637

LM num: 24
Nota =
-0.0067 * Amostra
- 0 * Avaliador
+ 0.007 * CodVinho
+ 0.3822 * AcidezvolatilgLacacet
- 0.0163 * pH
+ 0.0971 * AlcoolInfvv
+ 0.0006 * SO2livremgL
+ 0.0014 * SO2totalmgL
+ 1.9007 * SO2molecmgL
- 14.198 * DensidadegL
+ 0.044 * AcideztotalgLactart
- 0.3813 * ABS420nm
- 0.0002 * IPT
+ 0.0054 * ExtractosecogL
- 0.0004 * ExtractonaoredgL
- 0.0011 * Substanciasred
+ 0.0471 * FemgL
+ 0.1077 * CumgL
+ 26.6877

Anexo B4 (cont.)

LM num: 25
Nota =
-0.0043 * Amostra
- 0 * Avaliador
+ 0.0047 * CodVinho
+ 17.1888 * AcidezvolatilgLacacet
- 0.0163 * pH
+ 0.0971 * AlcoolInfvv
+ 0.0006 * SO2livremgL
+ 0.0014 * SO2totalmgL
+ 1.1898 * SO2molecmgL
- 14.198 * DensidadegL
+ 0.044 * AcideztotalgLactart
- 0.3813 * ABS420nm
- 0.0002 * IPT
+ 0.0054 * ExtractosecogL
- 0.0004 * ExtractonaoredgL
- 0.0011 * Substanciasred
+ 0.0471 * FemgL
+ 0.1077 * CumgL
+ 19.5704

LM num: 26
Nota =
0.0178 * Amostra
- 0 * Avaliador
+ 0.0005 * CodVinho
+ 0.3822 * AcidezvolatilgLacacet
- 0.0163 * pH
+ 1.367 * AlcoolInfvv
+ 0.0006 * SO2livremgL
+ 0.001 * SO2totalmgL
+ 0.1759 * SO2molecmgL
- 14.198 * DensidadegL
+ 0.249 * AcideztotalgLactart
- 0.3813 * ABS420nm
- 0.155 * IPT
+ 0.0054 * ExtractosecogL
- 0.0004 * ExtractonaoredgL
- 0.0011 * Substanciasred
+ 0.5836 * FemgL
+ 0.1077 * CumgL
+ 5.7535

Number of Rules : 26